**The Relationship between Online and Offline Measures of Gradient Sentence Acceptability**

by

Tamarae Hildebrandt

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Linguistics)
in the University of Michigan
2023

Doctoral Committee:

Associate Professor Jonathan R. Brennan, Chair
Professor Marlyse Baptista
Professor Julie E. Boland
Professor Acrisio Pires

Tamarae Hildebrandt

tamhil@umich.edu

ORCID iD: 0000-0002-2566-1222

**Acknowledgements**

I owe a lot to my various support systems from back home in Faribault, during my undergraduate career, and throughout the last 8 years in the Linguistics Department. I'm grateful for the support in helping me complete this dissertation.

*To my committee,*

Jon: thank you so much for all your advice, guidance, and support since the start of this project. I had little experience in conducting experiments, and I continually learned new skills. I appreciate your patience as I adjusted to being a parent and to have more realistic goals.

Marlyse, Acrisio, Julie: thank you all so much for your feedback, suggestions, and discussions on the dissertation over the past few years. I appreciate your support and patience as my deadlines shifted and experiments changed.

*To my family,*

Thank you to my parents, Todd and Angie, for listening to me talk about linguistics over an entire decade, explaining what I do to your friends, cheering my successes on, and letting me vent about stressful times whenever I call. Thank you to Christen, Kelby, Slade, Grandma Judy, Mimi & Harry, and Papa & Lois: You all have supported me through this dissertation and provided me the opportunity to set away from the stress. The time I spent with you helped recenter me, so I could focus on achieving my goals.

*To Jojo,*

Remember Spring term in 2011 when we were in the basement of Goodsell. Us, along with our newly established cohort, declared linguistics as a major one by one. I would have never guessed that would have been the start to both our linguistics career and our friendship. Thank you for pushing me to go outside of my comfort zone, challenging my beliefs about the world, and providing me unwavering support. From writing comps, working for AmeriCorps, moving to DC, moving to Michigan, earning our respective degrees, visiting each other, and simply just being a much-needed support system over the past decade.

*To my colleagues in the Linguistics Department,*

I owe a lot of thanks to the Linguistics Department over the years. You have given me many opportunities to travel, collaborate, and grow. I want to thank my fellow graduate students Dom Bouavichith, Ian Calloway, Batia Snir, Dave Odgen, Yourdanis Sedarous, Felicia Bisnath, Carrie Ann Morgan, Alicia Stevers, Dominique Canning, and Danielle Burgess for so many lounge conversations, friendships, watching bachelor, and game nights. Thank you to Ruaridh Purse and Emily Atkinson—as post-docs you all gave sound advice for the dissertation process, collaboration, and being good friends. Lastly, thank you so much to Pam Beddor and Jen Nguyen. You both move mountains to make things happen and I appreciate how fiercely you advocate for all of the graduate students—including me.

*To the outlets that kept me sane,*

I owe a lot to different activities that got me away from the dissertation when I needed it the most. Trivia is arguably the most important activity throughout my graduate school years. Over the past few years, the sense of community at Bløm Meadworks during trivia with "The First Place Team" and "Quaranteam" kept me coming back. The numerous D&D sessions that started during the height of COVID with the Feywild and Östime DND crews. Jojo, Justin, Dena, Ben, and Kell, you all let me escape and be silly by playing a Tabaxi Monk and a Warfordged Warlock. To my hometown friends: Desiree and Candis, thank you for the support and love through the years. To my Carleton friends: Becca, Elaine, Jojo, Mia, and Sarah, thank you for making me laugh with Zoom chats for an hour every month. The support we show each other, while we pursue our goals across the country, makes me exceptionally grateful.

*To Cati Fortin & Cherlon Ussery,*

Thank you both for fostering my interest in linguistics, for providing honest and sound advice, and for guiding me throughout the years. You've been there from term papers, to comps, to graduate school applications, to check-ins over the years, and now the dissertation. This would not have been possible without you both!

*To Sam Epstein,*

You are sorely missed. I couldn't have written this dissertation without your advice and guidance on my early projects in graduate school. Your insight and advice both academically and personally have remained with me. Two specific phrases got me through the dissertation: 1) how am I/ did I define this term clearly and 2) hang in there.

*And finally, to my buddies, Justin and Auggie,*

I'm so incredibly lucky to have you both here by my side. Thank you for the smiles and laughs when my stress was at the highest and sleep was at the lowest. Justin, thank you for your unwavering support through all facets of my life. I don't know how I got so lucky.

## Table of Contents

## List of Tables

## List of Figures

## List of Appendices

## Abstract

Experiments using electroencephalography (EEG) and self-paced reading (SPR) usually collect online and offline measures in separate tasks and typically at different points. Current research in psycholinguistic and neurolinguistic has separately shown that gradience exists in both offline acceptability judgments and online measures. These different types of measures are typically analyzed separately. And as such, analyzing these measures separately limits our understanding of the relationship between online and offline measures and how gradience affects the relationship. In this dissertation, I investigate whether gradience in offline acceptability judgments show proportional gradience in online measures.

This dissertation focuses on two target syntactic constructions: 1) a construction argued to exhibit gradient acceptability and 2) a dialectal construction from the US midlands. The first construction was examined in two experiments using two online measures, event-related potentials and reading times (Chapter 2 and Chapter 4). The second construction was examined in one experiment only using reading times (Chapter 5). In all three experiments, participants read sentences using the rapid serial visual presentation (RSVP) or the self-paced reading (SPR) protocols and the participants immediately rated the acceptability of the sentence on a 4-point Likert Scale, where 1 is *unacceptable* and 4 is *acceptable.*

The results of the experiments suggest that a slight negative correlation between the online and offline measures in SPR studies, but not in EEG. The EEG study only showed a reliable negative correlation for the subject-verb agreement construction and the SPR chapters (Chapter 4 and Chapter 5) showed a reliable negative correlation for all constructions.

Additionally, language experience based on geographical location did not modulate this negative correlation between online and offline measures in the dialectal construction using SPR via internet collection. This dissertation contributes to our understanding of how gradience within a syntactic construction, the type of online measure, and language experience impacts the relationship between online and offline measures.

# Chapter 1 Introduction

## 1.1 Motivation

Current psycholinguistic and neurolinguistic research has separately shown that gradience exists in both offline acceptability judgments and online measures collected from electroencephalography (EEG) and self-paced reading studies (e.g., Sprouse et al., 2013; Gouvea et al., 2010; Tanner, 2019; García, 2017; Zaharchuk et al., 2021; Squires, 2016). These studies analyze these online and offline measures separately as they were collected from a different experimental task at a different point within the experiment. This type of approach limits our understanding of the relationship between online and offline measures.

This dissertation develops a methodology to collect online and offline measures within the same trial and analyze their relationship. The overarching question of this dissertation is whether gradience in offline acceptability judgments (e.g., acceptability judgment responses that are distributed across the range of possible values) show proportional gradience in online measures. The dissertation includes a variety of online measures such as event-related potentials (see section 2.1.1) and self-paced reading times (see section 4.1 and 5.1) to allow for data collection in-person and via the internet. The dissertation focuses on two target constructions: 1) a construction argued to exhibit gradient acceptability (section 2.1.4.1) and 2) a dialectal construction (section 5.1.3). Offline acceptability judgments are the common measure used throughout this dissertation, which will be reviewed in section 1.2, I will then provide an overview of each chapter of the dissertation in section 1.3.

## 1.2 Offline Acceptability Judgments

Throughout this dissertation, I will use the term gradience in acceptability as judgment responses that are distributed across the range of possible values. This continuity view of acceptability is similar to Sprouse: "this continuity [acceptability of sentences is fairly evenly distributed across the range of possible acceptability levels], sometimes called gradience, appears to be a fact of acceptability judgments" (2015, p. 97).

### *1.2.1 Syntax and Acceptability Judgments*

An individual's mental grammar is called Internal Language (I-Language) and is defined as a part of an individual's mind that is acquired and used by the individual (Chomsky, 1986). The focus of Generative Syntactic Theory lies in representing the shared commonalities in grammar among individuals, as generative syntactic theories are constructed descriptively based on idealized speaker/hearer relationships (Chomsky, 2015) and how the forms and functions of sentences in various situations are used (Finegan, 2019). Variations in I-Language from the idealized form complicate our understanding of the underlying rules of the grammar.

The most common form of syntactic data used to construct, evaluate, and revise syntactic theories are acceptability judgments. These judgments are conscious reports from an individual about the degree of acceptability with which they perceived a stimulus or set of stimuli (Sprouse, 2022). Acceptability judgments are used as syntactic data because individuals cannot access grammaticality directly (Chomsky, 2015). Syntacticians use the term grammaticality to ascertain whether a sentence conforms to the rules of the grammar, while acceptability is the degree to which the <form *and* meaning> pair of a sentence is acceptable to speakers. These acceptability percepts are used as a proxy for grammatical well-formedness, thus providing the empirical foundation of many generative syntactic theories (Chomsky, 2015; Schütze, 2016).

### *1.2.2 Gradience in Acceptability*

Gradience has been examined under various rubrics including (1) gradience within the grammar itself, (2) task effects, (3) language experience (linguistic expertise, dialect etc.). To the first point, Sprouse (2015) broadly discusses two views that differ in where gradience exists. The first viewpoint is binary/categorical, and this idea aligns with the generative syntactic view above, in which the syntax either generates a sentence or it doesn't. The grammar, on this view, is not gradient and when gradience is present it is in non-syntactic cognitive systems. The second viewpoint consists of weighted-constraint theories, in which the syntax and non-syntactic cognitive systems all contribute to the acceptability of the sentence (see Sorace & Keller, 2005; Featherston, 2008; Bresnan, 2007; and Sprouse, 2015 for more information). However, the contribution of the syntax is higher than contributions from the non-syntactic cognitive systems. The various weights of the contributions lead to a range of acceptability judgments—more specifically gradience in acceptability. This dissertation looks at non-syntactic cognitive factors that affect acceptability judgments to see whether similar factors affect online implicit measures. Note that in this effort, I do not make a commitment as to whether gradience may also be directly encoded within the grammar. See Francis (2022) for more in-depth discussions on gradient acceptability and linguistic theory.

Gradience in acceptability has also been linked to experimental task effects like syntactic satiation. Syntactic satiation is when sentences that are first judged to be unacceptable are rated increasingly more acceptable over the course of an experiment (Snyder, 2000). Certain syntactic constructions might be susceptible to syntactic satiation effects within participants. For example Snyder (2000) observed significant satiation effects for whether-islands and complex NP constructions in English, but no significant satiation effects were observed for that-trace or left-

branching constructions. These results indicate acceptability ratings may change over the course of the experiment due to exposure. These changes present another pathway that leads to gradient acceptability judgments.

Another source of gradience in acceptability judgments may also arise from language experience, such as (1) linguistic expertise and (2) language backgrounds. Dąbrowska (2010) investigated whether acceptability judgments from expert linguists and naïve participants yielded similar results in grammatical sentences containing long-distance dependencies. The level of formal linguistic experience varied between these two groups of participants. In most cases, acceptability responses from the expert linguists and naïve participants diverged with expert linguists rating constructions as more acceptable. This was observed in a construction containing a prescriptive rule violation. Expert linguists rated the acceptability declarative sentences beginning with a conjunction (e.g., ***But*** *you think the witness will say something if they don't intervene*), as more acceptable than naïve participants. Sentences with prescriptive rule violations are not typically considered ungrammatical[1], but the violation clearly affect the acceptability judgment ratings of naïve participants. This study suggests that the experience level with linguistics affects intuitions about the acceptability of constructions.

Finally, languages users with dialectal experience may vary on how aware they are of their unconscious usage and their reported percept of acceptability. Labov (1996) has detailed several case studies (e.g., positive *anymore*, BIN in African American English, and usage of AIN'T) in which the linguistic behavior of participants conflicted with their intuitions. In one interview, Jack Greenberg provided strong intuitions that he did not use positive *anymore,* but a

[1] Crucially this assumes that an individual's I-Language contains rules allowing the prescriptive violations; however, some individual's I-Language could in fact contain rules in which prescriptive violation are also descriptively ungrammatical.

couple weeks later Jack was overheard saying *"Do you know what's a lousy show anymore? Johnny Carson* (Labov, 1996, p. 85)" to another co-worker while on a job. Jack clearly displayed a mismatch between his implicit knowledge of positive *anymore* and his lack of explicit awareness regarding his own usage of the same construction. This mismatch could result from social setting and/or social bias against this dialectal construction[2] leading to gradience in acceptability judgments.

Hildebrandt (2017) presented another example of language experience with a dialectal construction—preverbal *quick*[3]. Naturalistic data indicated that for some speakers the zero-form adverb *quick* is acceptable in the preverbal position: *I quick wrote an email to the professor*. But all speakers accept the preverbal adverb *quickly*: *I quickly wrote an email to the professor*. However, some speakers may not admit to using the preverbal *quick* construction outright, but the construction then comes out naturally at a later point in time—showing a mismatch in their language usage and their reported acceptability percepts. In an exploratory acceptability judgment task, Hildebrandt found that overall participants prefer *quickly* to *quick* in the preverbal position. However, geographical location influences how likely and individual is to accept preverbal *quick*, as native Michiganders and Minnesotans are more likely to accept quick in this position which could indicate that the internal grammar of these participants may allow for the preverbal *quick* construction.

To summarize, even under a framework with the grammar is discrete, these acceptability judgment studies have shown that these explicit measures may be gradient due to cognitive/processing factors and language experience. If a link between these offline measures

[2] At the time of the interviews, there was no known social stigma surrounded positive *anymore*, although change in social stigma may have occurred since the interviews were originally collected (Labov, 1996).
[3] I used several preverbal *quick* stimuli in Chapter 4 and Chapter 5 for the practice block. The construction is called Dialectal Adverb, as it includes two adverbs: *quickly* and *slowly*.

and online measures is expected, then the online measure may also show gradience due to the same factors as the offline measures. Section 2.1.2 will discuss cognitive factors affecting online measures using the P600. Section 5.1.1 and 5.1.2 will discuss evidence of language experience affects online measures including the P600 and RTs.

### 1.3 Overview of the Dissertation

The study in Chapter 2 tests whether gradient acceptability judgments to a target construction show proportional gradient neural responses to the P600. The target construction is sentences with variable complementizer realization (e.g., With: *The observation <u>that</u> six monkeys are swinging delighted the child*. Without: *The observation <u>these</u> six monkeys are swinging delighted the child*). As motivated in section 2.1.4.1, sentences without the complementizer have shown gradient acceptability judgments (Sprouse et al., 2013) but the complementizer construction (CP) has not been tested using EEG. In addition to the target construction, participants read and immediately rated the acceptability of each sentence across three other conditions: subject-verb agreement (SVA), gender reflexive (GR), and lexical semantic (SS). While the main effects of preference are not statistically reliable in the online measures for CP, GR, and SVA, the relationship between the online and offline measures was still examined. A linear mixed effects model was fit with amplitude as a dependent measure and acceptability, condition, and their interaction as a fixed effect. Results indicate that only the SVA condition shows statistically reliability in a negative correlation between the P600 amplitude and acceptability. Overall, these results do not support a proportional relationship between implicit neural and explicit acceptability judgment measures.

While the relationship between online and offline measures in Chapter 2 is not proportional, Chapter 3 explores whether the nonsignificance of the ERP signals could be

explained through gradience in the neural signal. Prior work suggests that individuals may differ, in a gradient way, in the balance between the relative strength of the N400 and P600 to a construction (Tanner, 2019; Tanner & Van Hell, 2014; Zeitlin, 2020). To assess the relative strength of these responses, the Response Dominance Index (defined in section 3.1) was calculated with EEG data from Chapter 2 for each participant in all four conditions within the N400 and P600 time frames. This study examines two key research questions: (1) do participants individually show positive-going RDIs in violations that traditionally evoke a P600 response and (2) do individual participants show stable RDIs (i.e., maintain the same RDI) across other violations that traditionally evoke a P600 response? The results indicate that RDI is unstable across the conditions to elicit the predicted P600 response, as participants did not maintain the same RDI for these conditions. This suggests that participants vary in the cognitive processing strategies they deploy (e.g., semantic integration and/or syntactic reanalysis) within a single condition. The use of different cognitive strategies within and across different conditions that typically evokes a P600 response likely explains the individual differences in RDIs among participants, as well as the unstable RDIs across the same conditions. Additionally, the individual differences in RDI may account for the gradience in the online measure in Chapter 2.

While Chapter 2 did not find a proportional relationship between the ERP amplitude and acceptability judgments, Chapter 4 explores the possibility of whether gradient acceptability judgments to a target condition show proportional gradience in online RTs using a self-paced reading task via the internet because of the COVID19 pandemic. Participants read and immediately rated the acceptability of each sentence across four conditions: complementizer, subject-verb agreement, gender reflexive, and lexical semantic. A linear mixed effects model was fit with RT as a dependent measure and acceptability, condition, and their interaction as a

fixed effect. The results show a small, but statistically reliable negative correlation between RTs and acceptability judgments across all four conditions. The overall results suggest a proportional relationship between offline and online measures across all conditions, in that as acceptability judgments increase, RTs decrease.

The study in Chapter 5 probes whether language experience with a dialectal condition modulates the negative correlation between acceptability judgments and RTs established in Chapter 4. The needs+past participle condition is likely to exhibit gradience in the acceptability judgment responses based on location of the participants. Therefore, the participants were recruited in two groups based on language experience based on current location and location born. These two groups read and immediately rated the acceptability of each sentence across four conditions: needs, phrase structure, gender reflexive, and lexical semantic. A linear mixed effects model was fit with RT as a dependent measure and acceptability, condition, group, and their interactions as a fixed effect. The results show that the proportional relationship in Chapter 4 is maintained when modulated by language experience. The results show a small, but statistically reliable negative correlation between RTs and acceptability judgments across all four conditions and both groups, in that as acceptability judgments increase, RTs decrease. Overall, regardless of language experience, a proportional relationship between offline and online measures was found within the dialectal condition suggesting that language experience affects offline and online measures similarly when collected within a single trial.

In Chapter 6, I summarize the contributions of this dissertation and explore future research directions, such as returning to EEG and assessing how awareness and sociolinguistic indexing may modulate the negative correlation between online RTs and offline acceptability judgments.

## Chapter 2 Exploring Gradience in Implicit Neural Measures and Explicit Acceptability Measures Using Electroencephalography (EEG)

Current psycholinguistic and neurolinguistic research has separately shown that gradience exists in both explicit acceptability judgments and implicit neurolinguistic responses collected from electroencephalography (EEG) (e.g., Sprouse et al., 2013; Gouvea et al., 2010). Typically, these two data types are collected and analyzed separately (Gouvea et al., 2010; García, 2017; Tanner, 2019); limiting efforts to understand how these measures might relate to each other. This experiment aims to explore the link between explicit acceptability judgments and implicit neural responses in a task targeting sentence acceptability. Specifically: Do gradient acceptability judgments to a target construction show proportional gradient neural responses to the P600? The two measures are compared in two linguistic conditions: i) sentences argued to exhibit gradient acceptability due to variable complementizer realization (Sprouse et al., 2013; Martin, 2001) and ii) subject-verb agreement sentences known to exhibit categorical acceptability (Tanner, 2018 & 2019). The later offers a sharper distinction between acceptable and unacceptable utterances to elicit a strong neural response and acceptability difference with which to contextualize the more tempered gradient differences expected for complementizer realization.

Lecky & Federmeier observe that "the P600 is a marker that can capture differences between the processing of syntactically congruent, preferred, or probable structures relative to incongruent, dispreferred, or less probable structures" (2019, p. 6). This perspective takes the P600 to reflect similar kinds of information as are typically attributed to offline acceptability

judgments. The latter have been argued to be impacted by a variety of different aspects of grammar or processing, including: “a syntactic constraint violation, a violation in a different part of the grammar, some component of sentence processing, word or construction frequency, sentence plausibility, or any number of other factors that impact sentence comprehension” (Sprouse, 2022, p. 10). Given these observations that similar cognitive and linguistic factors can modulate both acceptability judgments and the P600, the present experiment aims to better characterize if gradient acceptability is proportional to gradient EEG signals for specific syntactic constructions.

## 2.1 Literature Review

Recalling the literature reviewed in Chapter 1, even under a framework where the grammar is discrete, acceptability judgment studies have shown that these explicit measures may be gradient due to cognitive/processing factors and language experience. If a link between the explicit and implicit measures is expected, then the implicit measure may also show gradience to the implicit measure due to the same factors. Section 2.1.1 will discuss relevant language-specific event-related potentials used for the implicit measures within this experiment. Section 2.1.2 will briefly review how the implicit measures of human linguistic processing are connected to the grammar and will provide examples that the implicit signal (P600) can be gradient. Section 2.1.3 will explain the motivation for the research question and section 2.1.4 motivates the specific conditions used to target the research question.

### *2.1.1 Relevant Language-Specific Event-Related Potentials*

Electroencephalography (EEG) is an experimental methodology that uses electrodes to measure how electrical potentials on the scalp or cortex change over time (Luck, 2014; Swaab et

al., 2011). The electrodes reflect the summative activity of tens of thousands of cortical neurons (Coulson et al., 1998). When the raw EEG signal is then transformed by time and phase-locking the signal to an event (e.g., a semantically anomalous word in a sentence), this is called an event-related potential (ERP). ERPs are used to analyze whether systematic electrophysiological brain responses are present to a specific event measured in terms of four properties: amplitude, polarity, latency, and scalp topography (Luck, 2014). Two relevant language-specific ERPs are summarized in Table 2-1, giving for each component, type of deflection, onset of the component, and what linguistic events elicit the component response.

*Table 2-1: Relevant Language-Specific ERPs in Sentence Processing.* This work uses the N400 as a sanity check and the P600 to explore the research question at hand.

| ERP Components | Polarity and Topography | Latency of Component Post-Stimulus | What elicits the component? | Notable Citations |
|---|---|---|---|---|
| N400 | Negative-going; central-parietal region | 300-500 ms | All content words. Difficulty integrating semantic information (with context). Activation of semantic information in the lexicon. Discourse level manipulations of semantic features/information. | Kutas & Federmeier (2011); Kutas & Hillyard (1980) |
| P600 | Positive-going; central-parietal region | 500-800 ms | Phrase structure, agreement, tense, case, subjacency, and verb subcategorization violations. Garden-path sentences. Grammatical sentences that syntactically complex. Ambiguous sentences. Wh-movement. Island violations. Unexpected theta-role assignment. | Osterhout (1992) |

The N400 is a central-parietal negative-going voltage wave that occurs 300-500 ms post stimulus onset and peaks roughly at 400 ms (Kutas & Federmeier, 2011). All content words elicit the N400 to some degree; however, the N400 is especially sensitive the semantic integration and cloze probabilities of words in sentences and contexts (Kutas & Hillyard, 1980, 1984; Van Berkum et al., 2005, 2008). Select conditions and example sentences from Kutas & Hillyard (1980) are listed in (2-1).

*(2-1) Conditions and Example Sentences from Kutas & Hillyard (1980)*

a. Semantically congruent: I spread the warm bread with **butter.**
b. Semantically incongruent: I spread the warm bread with **socks.**

Participants show a smaller N400 response on the bolded word *butter* in (2-1)a because it is highly predictable and congruent with the semantic context of the sentence. In contrast, the bolded word *socks* in (2-1)b is highly unpredictable given the semantic context of the sentence. A larger N400 response is elicited when the semantic content of the target word is incongruent with the given sentential context.

Violations that elicit the P600 component include morphosyntactic violations, well-formed but syntactically-complex phrases, or violations in musical structure, mathematical rules, and other kinds of abstract sequences (Swaab et al., 2011; Tanner et al., 2018). The "P600" component occurs 500-800 ms after a violation and peaks around 600 ms with an increased positive voltage potential over the central-parietal region of the scalp (Swaab et al., 2011). Table 2-2 lists several linguistic violations and syntactically complex constructions that elicit the P600 component[4].

[4] The P600 response is commonly referred to as a single response type in the literature, but Lecky & Federmeier (2019) discuss two separate P600s: the syntactic P600 and the semantic P600. The syntactic P600 responds to morphosyntactic violations and grammatical structures that are difficult to process. In contrast, the semantic P600 is a response to some types of semantic anomalies (Kuperberg et al., 2003, 2007). It is hard to distinguish between the syntactic and semantic P600 response, as both elicit a positivity with a similar latency and scalp topography.

*Table 2-2: Types of Linguistic Constructions that Elicit the P600 Component.* Typically, these types of violations will elicit a larger P600 response relative to control sentences without that violation. Grammatically and syntactically complex sentences will also elicit a larger P600 response than control sentences without the grammatically complex part (e.g., long-distance dependencies).

| Constructions | Example Sentence or Type of Construction | Studies or Review Papers |
|---|---|---|
| Phrase Structure Violations | *We drank Lisa's by brandy the fire in the lobby.* | Neville et al. (1991); Osterhout & Holcomb (1992); Friederici and Meyer (2004); Davidson & Indefrey (2007); Batterink & Neville (2013) |
| Agreement Violations [Some examples: Number, gender, case, Tense] | *The cats will not eats the food.* | Osterhout & Holcomb (1992); Friederici and Meyer (2004); Gouvea et al. (2010); Tanner & Van Hell (2014); Tanner (2019) |
| Subjacency Violations | *I wonder which of his staff members$_i$ the candidate was annoyed when his son was questioned by ___$_i$.* | McKinnon (1996);Swaab et al. (2011) |
| Verb Subcategorization Violations | *Jill **entrusted** the recipe friends before she suddenly disappeared.* | Ainsworth-Darnell et al. (1998); Tanner et al. (2018) |
| Garden Path Effects | *The patient met the doctor **and** the nurse with the white dress showed the chart during the meeting.* | Gouvea et al. (2010); Tanner et al. (2018) |
| Animacy and Thematic Violations | *To make good documentaries cameras must interview…* | Kuperberg et al. (2007); Kuperberg et al. (2003) |
| Well-formed and Syntactically-Complex | Long-distance dependencies, Wh-questions | Swaab et al. (2011); Tanner et al. (2018) |

### *2.1.2 Gradience in the P600*

As shown in Table 2-2, constructions containing an outright syntactic violation or constructions that violate syntactic preferences and/or syntactic expectations are likely to evoke a P600 response. This response may be gradient in terms of the topography and the amplitude of the P600 response. Outright syntactic violations tend to show a posterior distribution of the P600 response, whereas violations of syntactic preferences (e.g., well-formed and syntactically complex) tend to show a more frontal distribution of the P600 response (Swaab et al., 2011).

With respect to amplitude, Gouvea et al. (2010) evaluated the internal structure of the P600 by presenting a within-subjects comparison of i) ungrammaticalities and ii) syntactic

garden-paths (e.g., well-formed but syntactically complex) in English. Both ungrammatical (agreement and phrase structure violations) and garden-path sentences elicited a P600 response, but ungrammatical sentences elicited a larger P600 than the syntactic garden-paths. This study provides evidence for a link between the processing of ungrammatical sentences and well-formed, if unexpected, sentences, in which ungrammatical sentences elicit a larger P600 response than well-formed sentences that also show a P600 response. If syntactic constructions vary in the strength of the P600 response, then this may also occur in the offline measure as well-formed sentences are likely to be rated as acceptable (depending on the complexity and preference of the sentence).

Table 2-3 summarizes a variety of studies that show graded P600 responses due to outright syntactic violations or violations of syntactic preference and/or expectations. I will highlight Dröge et. al (2016), and Osterhout et al. (1994) to demonstrate how violations in syntactic preference and syntactic expectation leads to a graded P600 response.

*Table 2-3: Graded P600 Responses*. (See also Molinaro et al. (2011) for a review of grammatical agreement processing in visual ERPs.)

| Authors | Cause of Graded P600 | Evidence for Gradience |
|---|---|---|
| Batterink & Neville (2013) [English] | Degree of Conscious Awareness | Detected violations when attention was reduced resulted in a smaller P600, whereas when attention was normal a higher P600. |
| Dröge et al. (2016) [German] | Salience of Word Orders using Morphosyntactic Violations | Grammatical, but dispreferred word order in the unambiguous case conditions resulted in a larger P600. |
| Hanulíková et al. (2012) [Dutch & Turkish-accented Dutch] | Expectations Regarding Speaker Identity | P600 is modulated by expectations of speaker identity (e.g., L2 accented speaker is expected to make mistakes therefore no P600 response). |
| Leinonen et al. (2008) [Finnish] | Degree Morphological Rule Violations | Combined inflectional and derivational violations elicited a larger P600 (a summation of the individual violations). |
| Mehravari et al. (2015) [English] | Degree of Morphological Violations | The interaction of morphological complexity and grammaticality modulated P600 amplitudes. |
| Nevins et al. (2007) [Hindi] | Salience and Degree of Agreement Violations | The salience feature (person) resulted in a larger P600 response than any other number, gender, and number/gender combination. |
| Osterhout et al. (1994) [English] | Frequency of a Subcategorization Complement | Statistical probability of sentential complement versus NP complement modulated P600 amplitudes. |

In an auditory ERP study, Dröge et. al (2016) investigated the processing of different word orders by varying case-marking and animacy in German. Broadly speaking, both the (preferred) subject-before-object (SO) and (dispreferred) object-before-subject (OS) word orders are grammatical; however, the OS word orders are marked. Dröge et al. predicted an attenuated P600 response to the unambiguous case-marked OS condition as compared to the unambiguous case-marked SO condition, as the case-marked nouns provided enough information for syntactic reanalysis. The results for the unambiguous case-marked OS condition indicated a large P600 response. This may indicate that the unambiguous case-marked nouns in the dispreferred OS condition caused more processing difficulty in reanalyzing the sentence, resulting in the higher P600 response, as opposed to an attenuated response.

A gradient P600 response was found in Osterhout et al. (1994) in sentences containing transitive or intransitive verbs. When participants syntactic expectations were violated to a greater degree (e.g., an outright violation of an intransitive verb taking an NP complement), then the largest P600 was elicited. An intermediate P600 was elicited with sentences that could take a dispreferred complement (either a sentential or an NP) of the transitive verb. If the participants syntactic expectations were met regarding the subcategorization requirements of the verb, then the P600 response was absent. This study suggests that when two syntactic options are possible, the dispreferred option may yield an attenuated P600 response.

I have separately argued that explicit acceptability judgments and implicit neural signals are gradient due to similar cognitive factors and language experience. Given this evidence, I next turn to examining these measures together to fully understand the relationship of these measures and further how gradience impacts the relationship. The following section examines the implicit and explicit data from García (2017) and describes the relationship between them in an effort to motivate the current study.

### *2.1.3 Research Question*

Research has shown that implicit ERP measures can be affected by sociolinguistic factors, such as linguistic background and language attitudes (Weissler & Brennan, 2020), just as explicit acceptability judgment responses are also affected by sociolinguistic factors (see section 1.2.2). As both implicit and explicit measures can be affected by a variety of cognitive, linguistic, and sociolinguistic factors, then it stands to reason that the EEG signals and acceptability judgments may reflect similar things, such as the syntactic knowledge of the participants. García (2017) has tested this theory of if there is a relationship between implicit ERP measures and acceptability judgments with mono-dialectal Mainstream United States

English (MUSE) speakers and bi-dialectal MUSE and African American English (AAE) speakers. She explored whether binary acceptability judgments (e.g., yes or no) reflected the P600 response using auditory stimuli from a bi-dialectal female speaker of MUSE and Southern American English[5]. Experimental stimuli and predictions are shown in Table 2-4 for both dialectal groups.

*Table 2-4: Condition and Group Design from García (2017, p. 47) with García's Predictions*

| **Conditions** | **Monodialectal MUSE Speakers** | **Bidialectal MUSE and AAE Speakers** |
|---|---|---|
| 3rd Person -S Marked<br>*The black cat laps the milk.* | Acceptable | Acceptable |
| 3rd Person -S Omitted<br>*The black cat lap the milk.* | Unacceptable | Acceptable, but MUSE-biased acceptability judgment predicted for participants |
| | P600 predicted | ***No P600 prediction made*** |
| Accusative Case (Agreement)<br>*The gentle doctor comforts them in the clinic.* | Acceptable | Acceptable |
| Nominative Case (Disagreement)<br>*The gentle doctor comforts they in the clinic.* | Unacceptable | Unacceptable |
| | P600 predicted | P600 predicted |

García predicted that the 3rd Person -S marked and accusative case conditions are acceptable for mono-dialectal speaker. The 3rd Person -S marked, 3rd Person -S Omitted, and accusative case conditions are acceptable for bi-dialectal speakers. Only the Nominative Case Disagreement was predicted to be unacceptable for both groups. While the 3rd Person -S Omitted condition is acceptable for bi-dialectal speakers, García predicted that acceptability judgments will be MUSE-biased and thus rate this sentence type as unacceptable[6]. The experiment lacked

[5] García states that Southern American English and AAE are dialects that have "well-documented feature overlap" (2017, p. 48). The speaker produced words from a word list with monopitch and monotone to control for intonation and inflection effects. García then used an auditory EPR study in which these words were presented one at a time with the duration of each word being one second. Given that visual and auditory modalities of presentation do not elicit significant differences in the P600 response (Balconi & Pozzoli, 2005; Hagoort & Brown, 2000; Hansen, 2005), this study serves as an acceptable predictor for what may occur in my experimental design.

[6] This is likely due to MUSE being the default dialect used in the United States and one that is commonly used in formal settings like experiments at a university.

an explicit prediction regarding whether the 3rd Person -S Omitted condition would elicit a P600 response in bi-dialectal speakers.

The implicit ERP measures and explicit acceptability judgment results for the Nominative Case Disagreement are listed below in Table 2-5. Both mono-dialectal speakers and bi-dialectal speakers rate the nominative case disagreement as unacceptable, and both groups show a P600 response to the case agreement violation. This suggests that implicit knowledge of ungrammatical constructions in MUSE and AAE were reflected in these language users' explicit judgments. In contrast to the Nominative Case Disagreement condition, Table 2-5 reveals a mismatch between the implicit ERP measure and the explicit acceptability judgments in bi-dialectal speakers for the 3rd Person -S Omitted condition. This result suggests that the implicit knowledge of a grammatical constructions is not always reflected in language users' explicit judgments when that construction is marginalized or minoritized (see also Zaharchuk et al., 2021 disscused in chapter 5).

*Table 2-5: García (2017) results to the sentence types of interest.*

| **Speaker Groups** | **Condition** | **Dialect** | **Acceptability Result** | **ERP Waveform Result** |
|---|---|---|---|---|
| Mono-dialectal MUSE | Nominative Case Disagreement | AAE and MUSE | 98.1% Unacceptable (Participant Response: No) | Evoked P600 Response (Positive deflection ~600 ms) |
| Bi-dialectal AAE and MUSE | Nominative Case Disagreement | | 91.8% Unacceptable (Participant Response: No) | Evoked P600 Response (Positive deflection ~600 ms) |
| Mono-dialectal MUSE | 3rd Person -S Omitted | AAE Only | 92.8% Unacceptable (Participant Response: No) | Evoked P600 Response (Positive deflection ~600 ms) |
| Bi-dialectal AAE and MUSE | 3rd Person -S Omitted | | 74.2% Unacceptable (Participant Response: No) | No evoked P600 Response (No positive deflection ~600 ms) |

García (2017) demonstrated that bi-dialectal speakers have knowledge of two grammars for their dialects (MUSE and AAE). Mono- and bi-dialectal groups lacked a P600 response to the nominative case violation of MUSE and rated this sentence type as unacceptable—meaning this construction violated a syntactic rule of MUSE. The bi-dialectal group lack a P600 response to the 3rd Person -S Omitted condition, which is grammatical in AAE and ungrammatical in MUSE. This indicated that the AAE dialectal construction was processed via rules from their AAE grammar. The difference waveform of the 3rd Person -S Omitted condition was nonsignificant and suggests no gradation due the omitted -S constructions. The P600 was not modulated by social factors in the bi-dialectal speakers and the acceptability judgments were gradient. The results suggest that the acceptability judgments were modulated by social factors in the bi-dialectal group.

One takeaway from García's work is that the relationship between the implicit and explicit measures is complicated. When a violation was detected in the Nominative Case Disagreement, participants across both groups showed a P600 response with unacceptable acceptability judgments. However, when a violation was detected in the 3rd Person -S Omitted condition, the linguistic experience of the bi-dialectal group affects this relationship between the implicit and explicit measures. The bi-dialectal group showed a mismatch in that there was not a P600 response and there were 74.2% unacceptable responses. This mismatch contrasts with the prediction that the lack of a P600 response would correlate with more acceptable responses.

This mismatch in online and offline measures observed by García and another mismatch observed by Zaharchuk et al. (2021) both lead to an important question about what the relationship between implicit and explicit measures are when both measures may be sensitive to different factors in gradience. My work recognizes that explicit acceptability judgments and

implicit neural signal vary gradiently not discretely, due to non-linguistic cognitive factors and sociolinguistic factors. It then aims to probe directly how implicit and explicit measures do, and do not, co-vary as indices of grammatical processing. To do this, I developed a methodology collecting both types of data within a single trial. I use a four-point Likert-scale to gather the acceptability judgments along a gradient. In terms of the gradience in the implicit ERP measure, grand average ERP waveforms do well in comparing how groups of participants perform on distinct conditions, but these waveforms cannot simultaneously examine the relationship between the amplitude of the P600 and the acceptability judgment rating or individual differences in implicit syntactic processes[7]. To simultaneously visualize these relationships, I add an alternative statistical analysis called ERP-Image Plotting (Delorme et al., 2015; Sassenhagen et al., 2014). ERP-Image Plotting is a more direct assessment of the relationship between the implicit neural signal and explicit acceptability judgment ratings with *syntactic constructions exhibiting gradience*, be it from personal preference or socially marked dialectal constructions.

Together, these changes to the experimental procedure and data visualization will better characterize what the relationship between explicit acceptability judgments and implicit ERP brain responses is, and how gradience among other factors may modulate these measures and the relationship between the explicit and implicit measures.

### *2.1.4 Stimuli & Conditions*

#### 2.1.4.1 Complementizer

A nominal complementizer condition was developed because the syntactic account of the null complementizer has been previously observed to elicit variable judgments. Martin (2001)

[7] These individual differences in implicit syntactic processes would be lost once their data is included in the grand average ERP waveform with the complete group of participants.

discusses the grammaticality[8] of the non-finite and finite clauses using the overt and null *for* and complementizer *that* in Table 2-6.

*Table 2-6: Non-finite for clauses versus finite that clauses Martin (2001; ex. 67 and 69)*

| ***for*** | **Complementizer *that*** |
|---|---|
| a. my desire [for my friend to win] | c. My belief that Kim is clever (is sincere). |
| b. *my desire [$Ø_{for}$ my friend to win] | d. *My belief $Ø_{comp}$ Kim is clever (is sincere). |

The overt *for* and *that* in Table 2-6(a and c) result in grammatical phrases. Null *for* in a non-finite clause results in an ungrammatical phrase in Table 2-6(b) and null *that* in a finite clause results in an ungrammatical sentence in Table 2-6(d). However, Martin (2001) argues that the degree of ungrammaticality of infinitivals is stronger than the degree of ungrammaticality of the finite clauses (fn 37). Martin suggests that finite clauses with the null complementizer *that* may reflect one of two possible syntactic analyses that could yield either a strong or weak violation: If the empty complementizer *that* fails to affix to the subject noun, then this leaves a stray affix within the syntactic derivation which, in turn, leads to a strong violation and presumably an outright unacceptable sentence. This analysis does not allow for gradience in the acceptability judgments and aligns with the binary/categorical view (section 1.2.2). On the other hand, if lexical *that* can be inserted in the syntactic derivation and then deleted at the PF interface, then this would lead to a weak violation. The deletion at PF may cause processing difficulties in perceivers and perceivers may develop a syntactic preference to keep or delete lexical *that,* perhaps based exposure to the construction, leading to gradience in acceptability of the weak violation.

[8] Martin uses the term grammaticality with these constructions, but I argue that the predicted grammaticality surrounding these constructions need further examination, following Section 1.2. The following syntactic proposals are discussed in terms of possible reported acceptability judgments to determine the role of gradience in each proposal.

These two syntactic analyses are tested in Sprouse et al. (2013). They collected acceptability judgments on sentences from this complementizer construction as part of a larger assessment of the reliability and validity of informal acceptability judgments. To do so, Sprouse et al. created a new set of eight minimal pairs that was rated by 141 participants. The results of the acceptability judgment task indicate that sentences without the complementizer (Table 2-7) are rated on average less acceptable than the sentences with the complementizer.

*Table 2-7: Acceptability Judgment Results: Mean, Median, and Quartile Measures with 7-point Likert Scale.*

| Condition | Mean | Median | Min. | Quartile 1 | Quartile 3 | Max. |
|---|---|---|---|---|---|---|
| With Complementizer | 6 | 6 | 1 | 5 | 7 | 7 |
| No Complementizer | 4.2 | 5 | | 3 | 6 | |

Crucially, within the data, the token responses for the complementizer condition are concentrated between 5 and 7, while for the no complementizer condition the tokens are concentrated between 3 and 6. In fact, the average acceptability judgment of the no complementizer condition falls at the mid-point (4.2) of the 7-point Likert-Scale. This wider range of responses and overlapping distribution within sentences with the complementizer indicates gradience among participants in rating sentences without the complementizer on a 7-point Likert-Scale. The gradience in acceptability judgments for the no complementizer condition provides evidence for the weak violation syntactic analysis, but not for the strong violations analysis, where no gradience is predicted.

To my knowledge, this specific complementizer condition (Table 2-6 c and d) has not been tested in an EEG experiment and it is unclear what the implicit neural measures would do in response to stimuli. The strong and weak violation proposals may align with previous research on the P600 response; recall, for example, that Gouvea et al. (2010) found that ungrammatical sentences elicited a larger P600 than grammatical well-formed, but hard to process sentences. The syntactic proposals from Martin (2001) could be examined to see if the strong or weak

acceptability intuitions match the implicit syntactic processes. To provide support for the strong violation, the P600 amplitude to sentences without the complementizer would be of a similar magnitude of a known ungrammatical violation (e.g., subject-verb disagreement). If the P600's amplitude of sentences without the complementizer was significantly smaller than a known ungrammatical violation, this would provide support for one of two possibilities: (1) either the sentence is grammatically well-formed, but difficult to process or (2) the sentence is grammatical but dispreferred leading to processing difficulty. However, the ERP results would not be able to indicate if this specific syntactic proposal is correct. Using finite sentences with and without complementizer *that* in this EEG study will provide neural evidence on the implicit syntactic processes of this condition.

#### 2.1.4.2 Subject-Verb Agreement, Gender Reflexive, and Lexical Semantic

Subject-verb agreement, gender reflexive, and the lexical semantic are deployed in this experiment along-side the complement condition just described. These stimuli are adapted from Tanner (2019) and are already controlled for word length, frequency, and concreteness. The stimuli are publicly available and were downloaded from the Harvard Dataverse (Tanner, 2018).

The subject-verb agreement condition is used as a clearcut syntactic violation in which the 3-person singular subject disagrees in the number feature with the lexical verb. The gender reflexive condition is a clearcut pronominal violation based on language stereotypes in which the reflexive disagrees in the gender feature of the subject. Both conditions are expected to yield a "Canonical" P600 response. The lexical semantic condition is included to complement the syntactic violation; it presents a clearcut lexical semantic violation in which a lexical item conflicts with the sentential context. As discussed in 2.1.1, when the lexical item conflicts with the sentential context, then this incongruity should elicit an N400 effect. This ensures the EEG

signal is operating as predicted. Examples of each condition with the congruous and incongruous preferences are shown in section 2.2.3 in Table 2-8.

## 2.2 Methods

### *2.2.1 Participants*

Twenty-nine participants[9] (21F, 8 M) were recruited for the study that took place in the Computational Neurolinguistics EEG Lab at the University of Michigan. Participants were 18 years of age or older (mean age 22.2), spoke English as a primary language before the age of 6, had no history of neurological disorders based on self-report, have normal or corrected-to-normal vision, and had a positive handedness score from the Handedness Survey from (Oldfield, 1971). At the conclusion of the study, participants completed a sociolinguistic demographic survey, which probed the participants' native and acquired languages, native and acquired dialects, age, gender, locations participants have lived, and general impressions about the experiment or experimental stimuli. More than half of all participants were born and raised in Michigan (17). Participants were compensated $15 per hour for their participation.

### *2.2.2 EEG and Acceptability Judgment Procedure*

The experiment took about an hour and a half on average to complete. The EEG data were recorded with 61 actively amplified electrodes and one ground electrode (ActiCap, Brain Products GmbH) using an elastic cap with the Easycap M10 layout. Electrode impedances were kept at 25kΩ or below. Data were recorded at 500 Hz with a hardware band-pass filter of 0.1 and 200 Hz. The electro-oculogram (EOG) was recorded from electrodes placed above and below the left eye.

[9] Data collection stopped before the intended goal of 30 participants due to hardware difficulties with the EEG.

The EEG experiment was deployed using PsychoPy (Peirce et al., 2019) while participants were seated in an isolated booth for the task. Before the experiment began, participants read an instructions page explaining the mechanics of the experiment and how to rate the acceptability of sentences (Figure 2-1):

"You will rate the acceptability of each sentence with a number: 1 unacceptable, 2 slightly unacceptable, 3 slightly acceptable, 4 acceptable. Consider two things when rating the acceptability of a sentence: 1) Is something you or another native speaker of English could say, or 2) Is something you could write in everyday life, not just in essays for class".

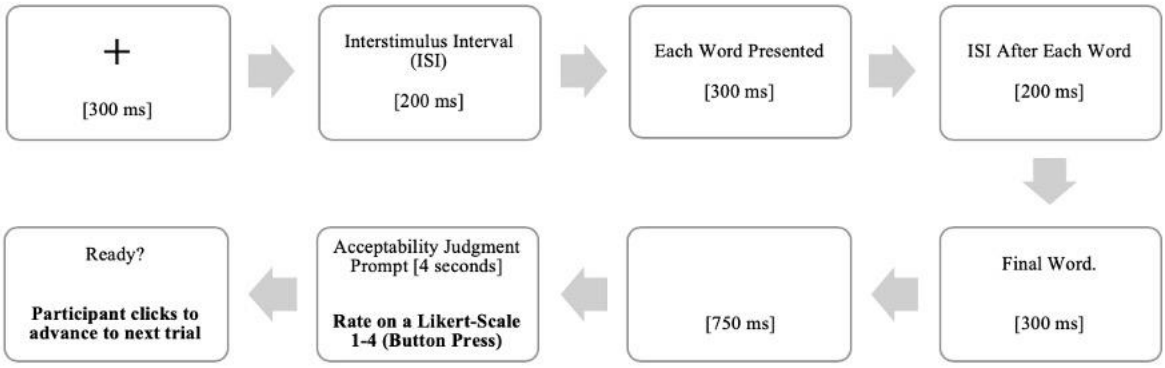


*Figure 2-1: Experimental Procedure using the Rapid Series Visual Presentation Paradigm.* The acceptability judgment scale is the key difference between previous EEG studies (Neville et al., 1991; Tanner, 2019) and this study. Most EEG studies use binary choices for sentence acceptability, whereas the acceptability judgment scale in this study used a 4-point Likert-scale (1 unacceptable to 4 acceptable) to rate the acceptability of the previous sentence.

As shown in Figure 2-1, each trial began with a fixation cross. Participants read sentences word-by-word (300 ms word presentation, 200 ms ISI) using the Rapid Serial Visual Presentation (RSVP) protocol (Gouvea et al., 2010; Hagoort & Brown, 2000; Neville et al., 1991; Tanner, 2019). The end of the sentence was indicated with a period. At the end of each trial, participants rated sentence acceptability using a four-point Likert-scale (Dröge et al., 2016). The acceptability judgment prompt remained on the screen for maximally 4 seconds or until the participant entered their response. Each trial ends with the "Ready?" screen for participants to prepare themselves for the next trial. Breaks follow each block of the experiment allowing

participants to rest their eyes, drink some water, and/or mentally reset before preceding to the next block.

This EEG procedure collects online and offline data within a single trial nearly simultaneous, which is similar to studies in the 1990s by Osterhout and colleagues (Osterhout, 1997; Osterhout et al., 1994, 1996, 1997; Osterhout & Holcomb, 1992; Osterhout & Mobley, 1995; Osterhout & Nicol, 1999). Two key differences between these Osterhout and colleagues studies and the procedure in Figure 2-1 are 1) following the final word with a period, a blank screen of 1450 ms and 2) participants answered a binary acceptability judgment (e.g., acceptable and unacceptable[10]) about the sentence in the trial. One disadvantage to collecting near simultaneous online and offline measures is the effect of task relevance, as thinking about the various anomalies and violations may bring them to the forefront of the participants attention, which in turn may elicit a P300 response.

A classic P300 response peaks between 300-400 ms (Leckey & Federmeier, 2019), although this positive broad positive going wave with a centro-parietal distribution can occur anytime between 250-1000+ ms after the critical stimulus onset (Polich, 2011). P300 responses typically follows highly salient events (e.g., novel or unexpected events), task relevant expected events (e.g., categorizing the stimuli), and subjective probability (e.g., how likely a stimulus belongs or is thought to belong to a category) (Leckey & Federmeier, 2019; Sassenhagen et al., 2014; Luck, 2014; Polich, 2011; Swaab et al., 2011). The events or subjective probability that elicits a P300 response can be linguistic or non-linguistic.

[10] Acceptable was defined as both semantically coherent and syntactically well-formed, as several studies used semantically anomalous sentences, verb tense violations, gender reflexive violations, and reduced relative clause stimuli (Osterhout, 1997; Osterhout et al., 1994, 1997; Osterhout & Holcomb, 1992; Osterhout & Mobley, 1995; Osterhout & Nicol, 1999).

Osterhout et al. (1996) examined if the syntactic P600 response is part of the P300 through testing a subject-verb agreement violation, a physical anomaly (e.g., all uppercase letters on the target word), and a double anomaly containing the physical difference and subject-verb disagreement. Osterhout et al. used a RSVP procedure (350 ms word presentation, 350 ms ISI) and ended every trial with a blank screen for 1450 ms followed by a binary acceptability judgment prompt. The results indicated that a P600 response was elicited for the agreement violation, a P300 response was elicited for the physical anomaly, and the double anomaly showed an additive effect (e.g., a P300 followed by a P600 response). This suggests that the P600 response "elicited by agreement violations is distinct from the P300 response to unexpected, task-relevant anomalies that do not involve the violation of a grammatical rule" (Osterhout et al., 1996, p. 507). As the P300 and P600 are distinct in their responses (e.g., time course, amplitude, scalp distribution) while using a similar EEG procedure[11], this Osterhout et al. experiment suggests that the task relevance of the acceptability judgments is unlikely to impact the neural response for the participants for the procedure in Figure 2-1.

The experimental design in Figure 2-1 makes use of an even-point Likert-scale, which is less common than using an odd-point Likert-scale. The even-point scale was selected to ensure that the relationship between the explicit acceptability judgment and the implicit neural measure would be interpretable from the ERP-image (discussed in section 2.2.5), as participants would not be able to use a mid-point to indicate neutrality or to avoid rating the sentence altogether.

[11] More work is needed identity hypothesis, which is the P600 as a member of the P3 family (Leckey & Federmeier, 2019; Dröge et al., 2016; Sassenhagen et al., 2014; Bornkessel-Schlesewsky et al., 2011). Some empirical support for the P600-as-P3-hypothesis comes from Coulson et al. (1998), who presented a visual ERP study to test the sensitivity of the P600 to the probability and frequency of grammatical errors (pronoun case disagreements and subject-verb disagreement). Their results indicated that P600 effects are sensitive to the salience and probability of these morphosyntactic violations; however, the P3b is also sensitive to salience and probability, and these results could not tease apart the difference between the P3b and the P600 response. The evidence suggests that these components may interact as they are elicited by similar factors within similar latencies. More work is needed to determine if the P600 is a member of the P3 family.

The 4-point scale also created ideal conditions for minimal movement, as participants were able to place their left fingers on numbers one through four without needing to move their entire hand. Thus, using a four-point Likert-scale results may provide cleaner data, more interpretable results, and a more nuanced understanding of if non-syntactic cognitive factors affect the implicit measures.

### *2.2.3 Stimuli Design*

520 sentence stimuli were divided into target (160), control (120), and filler (240) conditions—consisting of complementizer sentences, subject- verb agreement, lexical semantic sentences, and gender reflexive sentences respectively. Each of these conditions was presented with a congruous or incongruous variant: with/without the complementizer, grammatical/ ungrammatical subject-verb agreement, felicitous/infelicitous lexical semantic sentences, and grammatical/ungrammatical gender reflexive pronoun agreement.

*Table 2-8: Experimental Stimuli.* The underlined word is predicted to elicit either the P600 (Complementizer, Subject-Verb Agreement, and Gender Reflexive) or the N400 (Lexical Semantic) ERP components.

| Condition | Sentences Per Block | Example Stimuli |
|---|---|---|
| Target: With or Without Complementizer *that* [80 sentences] | 4 | *With Complementizer that [Congruous]*<br>The observation <u>that</u> six monkeys are swinging delighted the child.<br>The belief <u>that</u> seven baristas are coffee snobs is widely accepted. |
| | 4 | *Without Complementizer that [Incongruous]*[12]<br>The observation <u>these</u> six monkeys are swinging delighted the child.<br>The belief <u>these</u> seven baristas are coffee snobs is widely accepted. |
| Control: Subject-Verb Agreement [60 sentences] | *3* | *Grammatical [Congruous]*<br>The cats <u>meow</u> by the window watching the birds.<br>The ducks <u>swim</u> in the lake enjoying the spring breeze. |
| | *3* | *Ungrammatical [Incongruous]*<br>The cats <u>meows</u> by the window watching the birds.<br>The ducks <u>swims</u> in the lake enjoying the spring breeze. |
| Filler: Lexical Semantic [60 sentences] | *3* | *Felicitous [Congruous]*<br>The child borrowed some <u>books</u> from the library.<br>The gardeners trimmed the <u>shrubs</u> last Monday. |
| | *3* | *Infelicitous [Incongruous]*<br>The children borrowed some <u>conversations</u> from the library.<br>The gardeners trimmed the <u>purse</u> last Monday. |
| Filler: Gender Reflexive [60 sentences] | *3* | *Grammatical [Congruous]*<br>The calm bride prepared <u>herself</u> for the wedding.<br>The nervous groom checked <u>himself</u> in the mirror. |
| | *3* | *Ungrammatical [Incongruous]*<br>The calm bride prepared <u>himself</u> for the wedding.<br>The nervous groom checked <u>herself</u> in the mirror. |

The sentences were separated into two lists, such that each participant read and rated the acceptability of 260 American English sentences to avoid repetition effects. The experiment was

[12] The incongruous preference adds the demonstrative *these* and it was added to the stimuli to keep the stimuli as minimally different as possible; however, this addition may have affected the semantic interpretation of the incongruous stimuli, as discovered in post-experimental discussions regarding the construction and preferences.

split into 10 blocks each containing 26 sentences. Each block took approximately 4 minutes to complete with each block of sentences averaging about 10 words per sentence. As shown in Table 2-8 each block consisted of a fixed number of sentences: 8 from the Target and 6 sentences each from the Subject-Verb Agreement Controls, Lexical Semantic Fillers, and Gender Reflexive Fillers. As the target word's location varied across each condition (see underlined target words in Table 2-8), the average proportion of target word to sentence length in each condition varies from 0.3 to 0.63 (Table 2-9). The position of the target word as a proportion of the total sentence length in the Complementizer condition and the Subject-Verb Agreement condition were nearly identical.

*Table 2-9: Position of Target Word as a Proportion of Total Sentence Length for Stimuli in Each Condition.*

| Condition | Position of Target Word to Total Sentence Length | Average Sentence Length | Range of Sentence Length |
|---|---|---|---|
| Complementizer | 0.30 | 10 | 9 to 13 (4) |
| Subject-Verb Agreement | 0.32 | 11 | 6 to 15 (9) |
| Gender Reflexive | 0.63 | 9 | 7 to 12 (5) |
| Lexical Semantic | 0.57 | 10 | 7 to 16 (9) |

### *2.2.4 Hypotheses & Predictions*

Table 2-10 provides predictions for all four conditions within the EEG experiment. These include what kind of ERP response is elicited, the acceptability ratings that are expected, and what the relationship between the implicit and explicit measures might be (e.g., a relationship that operates in a proportional fashion).

*Table 2-10: Implicit and Explicit Measure Predictions in each Condition.* ERP Responses with ^ indicate the key condition where a lock-step relationship between the P600 response and acceptability is predicted.

| Condition | Grammaticality Stated from Previous Literature | General Acceptability Type | ERP Response | Likert-Scale Rating |
|---|---|---|---|---|
| Target: Complementizer *that* | Grammatical | Congruous | No P600 | 3-4 |
| | Ungrammatical | Incongruous | Variable Sized P600^ | Graded Responses 1-4 |
| Control: Subject-Verb Agreement | Grammatical | Congruous | No P600 | 3-4 |
| | Ungrammatical | Incongruous | Yes P600 | 1-2 |
| Gender Reflexive | Grammatical | Congruous | No P600 | 3-4 |
| | Ungrammatical | Incongruous | Yes P600 | 1-2 |
| Lexical Semantic | Grammatical | Congruous | No N400 | 3-4 |
| | Ungrammatical | Incongruous | Yes N400 | 1-2 |

#### 2.2.4.1 Predictions of the Implicit Neural Responses.

Each incongruous sentence type within this experiment (Table 2-10) is predicted to elicit a distinct language ERP component. The P600 response is predicted to occur for the control condition with subject-verb agreement violations (Tanner, 2019; Gouvea et al., 2010; Osterhout & Holcomb, 1992), but there is no evidence whether a P600 response will be elicited to the target condition without complementizer *that*. The gender reflexive condition is predicted to evoke a P600 response because the gender feature on the antecedent in the ungrammatical sentences does not match the gender feature of the subject—thereby causing a morphosyntactic gender feature violation (Osterhout et al., 1997). The lexical semantic condition will evoke an N400 effect in the incongruous sentences because the critical word conflicts with the

participant's expectation and/or integration given the sentential context. It is possible that the gender reflexive condition may also evoke an N400 effect for the same reasoning as the lexical semantic condition.

#### 2.2.4.2 Predictions of the Explicit Acceptability Judgment Responses.

Predictions regarding the explicit acceptability judgment response correspond to prior results in the literature and my own norming study results with these items. The incongruous complementizer condition is the only condition predicted to demonstrate gradience in judgments, as evidenced by Sprouse et al. (2013) and the norming results. The other three conditions are predicted to be rated categorically as acceptable or unacceptable based on the condition's grammaticality status. Congruous items (e.g., sentences with the complementizer and subject-verb agreement) will be rated categorically as acceptable (Likert-Scale rating 4). The incongruous subject-verb agreement is predicted to be rated categorically as unacceptable (Likert-Scale rating 1) because of the mismatch in the number features—causing a morphosyntactic violation.

#### 2.2.4.3 Predictions of the Relationship between the Implicit and Explicit Measures.

ERP-images (Delorme et al., 2015; Sassenhagen et al., 2014) are used to provide a fine-grain visualization of the relationship between the P600 response and acceptability judgment responses in each condition for all participants. If the relationship between the P600 response and acceptability judgment responses operates in a proportional fashion to a condition demonstrating gradient acceptability judgments, then we would expect to see the highest positive voltage with the unacceptable rating of 1 on the Likert-scale because a morphosyntactic violation is expected to evoke a higher positivity than a baseline condition. The intensity of the P600 response should diminish with a higher rating on the Likert-scale, indicating that there is less of a

positivity when the condition is less of a violation. If this prediction is realized, then when gradation is present in the explicit acceptability judgment, the implicit electrophysiological signal is also affected to the same degree.

### *2.2.5 Data Preprocessing & Analysis*

The raw EEG data was band-pass filtered between 0.5–40 Hz and divided into 1300 ms epochs around each target word in a sentence. Ocular signals were removed with ICA, and other artifacts were visually identified and excluded using Fieldtrip in MATLAB (*MATLAB*, 2019; Oostenveld et al., 2011). At the conclusion of data preprocessing, seven participants' data were excluded from the analysis due to noisy data (i.e., rejecting more than 20% of the 260 trials).

*Table 2-11: Summary of Trials and Data Exclusion.* The Accepted EEG Trials Column includes all trials used the visualizations of the EEG data.

| Condition | Type | Total Trials Possible | Accepted EEG Trials | Rejected EEG Trials | Number of Excluded No-Judgment Trials[13] | Percentage of All Excluded Trials from Data analysis |
|---|---|---|---|---|---|---|
| Complementizer | Congruous | 880 | 743 | 137 | 9 | 16.5% |
| | Incongruous | 880 | 754 | 126 | 10 | 15.5% |
| Subject-Verb Agreement | Congruous | 660 | 560 | 100 | 3 | 15.6% |
| | Incongruous | 660 | 551 | 109 | 4 | 17.1% |
| Lexical Semantic | Congruous | 660 | 577 | 83 | 6 | 13.5% |
| | Incongruous | 660 | 567 | 93 | 5 | 14.8% |
| Gender Reflexive | Congruous | 660 | 574 | 86 | 2 | 13.3% |
| | Incongruous | 660 | 575 | 85 | 4 | 13.5% |

Processed data is examined in two types of critical graphs: grand average ERP waveforms and ERP-images (e.g., Delorme et al., 2015; Sassenhagen et al., 2014). Grand average ERP amplitude is analyzed at 600ms for a P600 response; a response linked with a morphosyntactic violation in the complementizer, subject-verb agreement, and gender reflexive

[13] N/A judgment responses appear in the ERP-Images at the bottom of the black line closest to the x-axis (Figure 2-4).

conditions. Given that grand average ERP waveforms are unable to show gradation, ERP-images are also used which provide a fine-grain visualization of the relationship between acceptability judgments and the P600 response in each condition for all participants.

To assess the relationship between acceptability judgments and amplitude, I fit a linear mixed effects model with amplitude as a dependent measure and acceptability, condition, and their interaction as a fixed effect. I included subjects as a random intercept and random slopes for their acceptability, condition, and their interaction. Priors were set using the defaults of the brms package (version 2.17.0) in R.

## 2.3 Results

The results section will be structured by first examining these measures separately before the analysis of both measures simultaneously. First, I will begin with a discussion of the explicit acceptability judgment responses in the norming study and the EEG experiment. Next, the implicit electrophysiological measures will be examined. Then, the explicit and implicit measures will be combined into the grand average ERP-image to visualize the relationship between the explicit and implicit measures. The relationship between the implicit and explicit measures will be further examined through a linear mixed effects model.

### *2.3.1 Explicit Behavioral Acceptability Judgment Responses*

Figure 2-2 shows the acceptability judgment responses in the norming study (left) and the EEG experiment (right) for the target and control conditions (e.g., Complementizer and Subject-Verb Agreement). The norming study results (Figure 2-2; left) indicates that the congruous complementizer and congruous subject-verb agreement were rated categorically as acceptable (Likert-Scale rating 4), and the incongruous subject-verb agreement were rated categorically as

unacceptable (Likert-Scale rating 1). In contrast to these three categorically rated conditions, the incongruous complementizer condition elicits the by-subject average acceptability judgments at all four points in the Likert-Scale—yielding gradience.

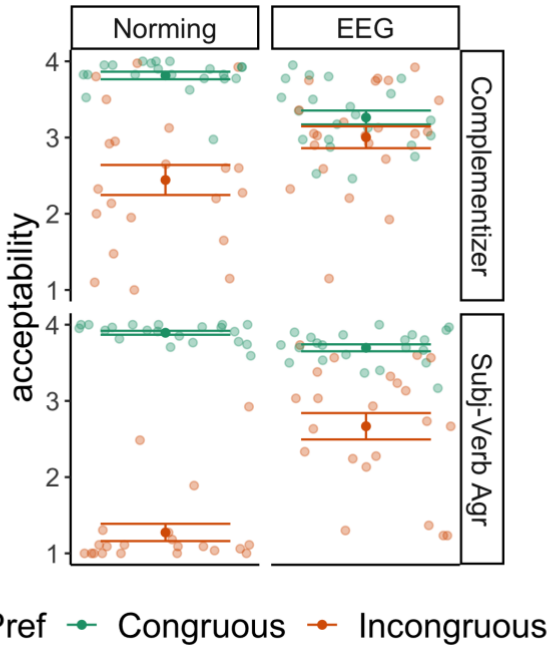


*Figure 2-2: Subject Average and Grand Acceptability Judgment Responses in Norming study and EEG experiment.* Participants rated the acceptability of stimuli on a 4-point Likert-Scale (1 unacceptable to 4 acceptable) in a Norming study (left panel) and EEG experiment (right panel). Stimuli consists of two conditions (Complementizer and Subject-Verb Agreement) across two preferences (Congruous and Incongruous). Subject averages are represented by transparent dots and the grand average acceptability and standard error are represented by the bolded dots and lines. [Norming study n=21; EEG experiment n=29].

The EEG experiment (Figure 2-2; right) results indicated gradience in the incongruous stimuli in both conditions. By-subject averages show acceptability judgment ratings across all four points of the Likert-Scale in the incongruous complementizer condition and the incongruous subject-verb agreement condition. This suggests that participants rated sentences with agreement feature mismatches more gradiently in the EEG experiment, but not the norming study, even though the incongruous subject-verb stimuli appear to be outright ungrammatical.

Figure 2-2 suggests that there is a stark difference in the ratings of the subject-verb agreement condition between the norming study and the EEG study. These studies varied in the

sentence presentation and length of time given to make an acceptability judgment. The norming study, which was conducted on Qualtrics, displayed ten sentences per page and the participants were able to read and rate the sentences at their own pace. As a time-limit was not enforced, participants could reread the sentences as many times as they needed before making an acceptability judgment. As a result, the reported acceptability judgments in the norming study were less noisy than the EEG study because participant did not need to rely on their memory of what they had read. The RSVP stimuli presentation in the EEG experiment, the inability to reread the sentences, and an enforced acceptability judgment time limit (e.g., maximum of 4000 ms) undoubtedly affected how participants remembered the ungrammatical subject-verb agreement condition and rated the condition.

The results from the other three conditions appear to behave as predicted, suggesting that the ratings collected in this experiment are reliable proxies for acceptability. Figure 2-3 presents the explicit acceptability data from the target complementizer condition, the subject-verb agreement condition, along-with the two other conditions used in the EEG experiment.

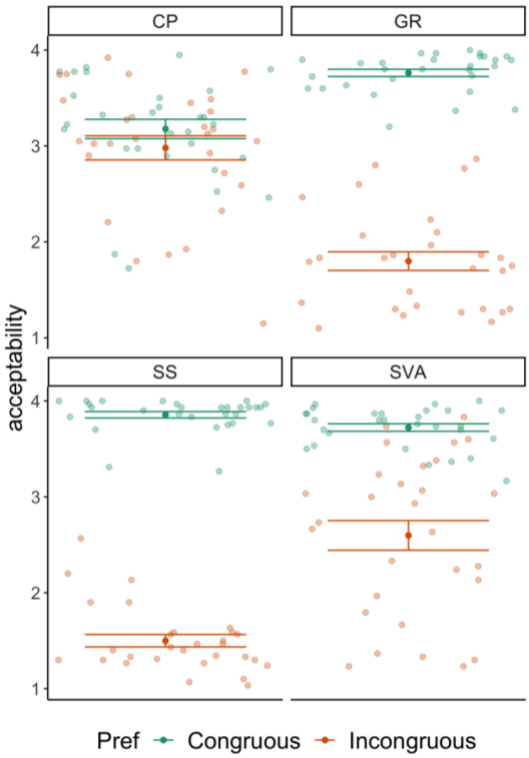


*Figure 2-3: Subject Average and Grand Acceptability Judgment Responses in the EEG experiment.* Participants rated the acceptability of stimuli on a 4-point Likert-Scale (1 unacceptable to 4 acceptable). Stimuli consists of four conditions (Complementizer, Subject-Verb Agreement, Gender Reflexive, and Lexical Semantic) and two preferences (Congruous and Incongruous). Subject averages are represented by transparent dots and the grand average acceptability and standard error are represented by the bolded dots and lines. The incongruous preference in the Complementizer and Subject-Verb Agreement conditions show gradation in acceptability with higher acceptability ratings, while the incongruous preference in the Gender Reflexive and Lexical Semantic show categoricity with the predicted unacceptable acceptability ratings. [n=22].

The congruous preferences (green) in Figure 2-3 follows the prediction in section 2.2.4.2 in that the average acceptability judgment is rated between 3 and 4. The incongruous preferences (orange) shows more gradience across these four conditions. The Gender Reflexive and Lexical Semantic conditions show the predicted average acceptability judgments fall between 1 and 2. Again, the average acceptability judgments for the Complementizer and Subject-Verb Agreement conditions are rated higher between 2 and 3 due to more gradience in the by-subject averages. The graded responses were expected for the incongruous Complementizer condition but were not expected for the incongruous Subject-Verb Agreement condition.

### *2.3.2 Implicit Neural Measures*

A non-parametric cluster analysis was conducted on the ERP data for all four conditions. The analysis was nonsignificant for the Complementizer condition, Subject-Verb Agreement condition, and the Gender Reflexive condition. The non-parametric cluster analysis did show a significant N400 effect in the lexical semantic condition. While the main effects of preference are not statistically reliable, I consider the quantitative patterns and differences between the conditions below.

Given the Gender Reflexive condition show the predicted explicit response, I selected the electrodes for data analysis based on the implicit measure in the Gender Reflexive condition; this showed a strong P600 effect for the electrodes Pz, CPz, Cz, CP3, CP1, and P1.

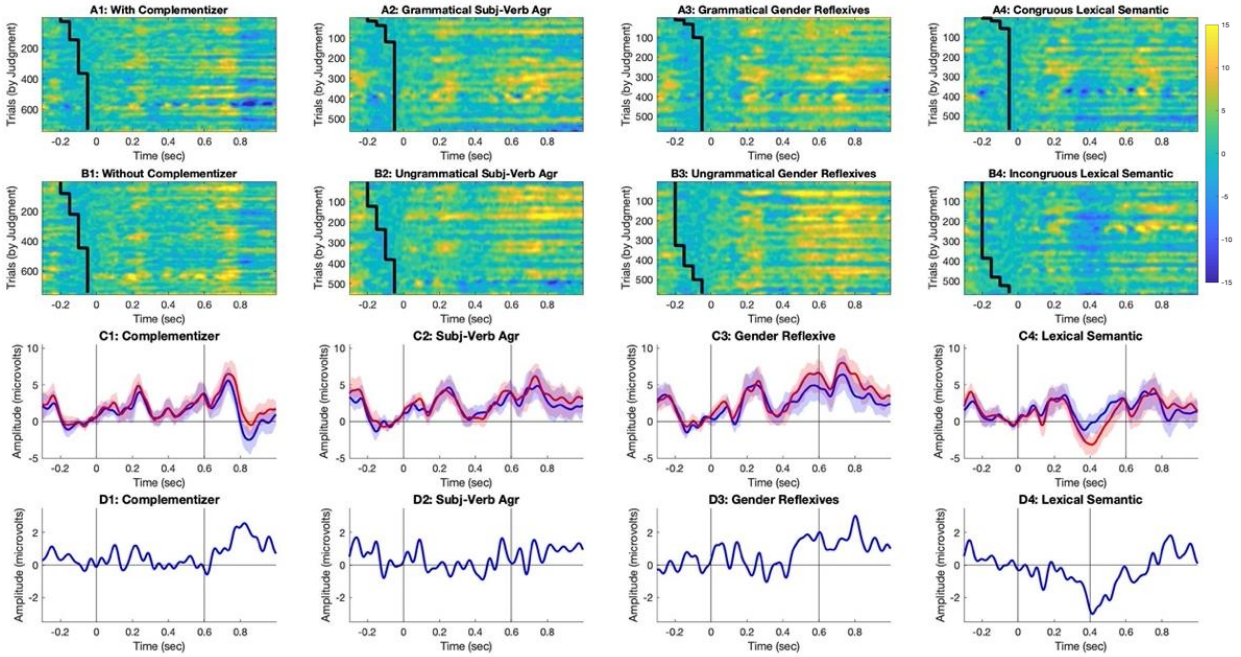


*Figure 2-4: Complete Summary of the EEG Results.* ERP-Images for the congruous sentence types are in the first row (A1-A4) and the incongruous sentence types are in the second row (B1-B4). Grand average ERP waveforms are in the third row (C1-C4), where the congruous preferences are shown in blue (e.g., predicted grammatical), and the incongruous preferences are shown in red (e.g., predicted ungrammatical). The difference waveforms in the fourth row (D1-D4) are calculated by subtracting the congruous from the incongruous sentence type. [Electrodes: Pz, CPz, Cz, CP3, CP1, P1. n=22.]

The ERP-Images across these electrodes (Figure 2-4- rows A and B) provides a way to look at the implicit neural data across time as a function of voltage amplitude. Time is represented along the x-axis, while the P600 response is measures in voltage and shown in color on the ERP-images[14]. Each row in the ERP-image is organized by the given acceptability response for each trial (black line from unacceptable (top) to acceptable (bottom)) on the y-axis. Figure 2-4 rows A and B show the congruous and incongruous conditions, respectively, for all four constructions (columns). The ERP-images show varying strengths of a P600 responses across all four-points of the acceptability judgment scale (e.g., see B3 the incongruous gender reflexive ERP-image). This indicates that participants have rated some of the incongruous preferences as acceptable, but their neural signal is trending towards a syntactic violation.

The Complementizer condition (left-most column) shows a stronger P600 response for sentences without the complementizer in comparison to the baseline with a complementizer. The subject-verb agreement condition trends toward a weak P600 response for sentences with incongruous subject-verb agreement in comparison to the congruous baseline. Note that this implicit neural response is consistent with the variable acceptability judgment responses recorded in the EEG experiment (Figure 2-2 & Figure 2-3). Figure 2-4(row C) shows the grand average ERP waveforms of the complementizer condition and the subject-verb agreement conditions. The congruous conditions (e.g., with the complementizer and grammatical Subj-Verb Agr) are represented by the blue lines, while the red line represents the incongruous conditions (e.g., without the complementizer and ungrammatical Subj-Verb Agr). Comparing the grand average Subject-Verb Agreement to the Complementizer condition shows that the participants

[14] The color scale, which goes from yellow (high positivity; 15), orange (medium positivity), green (low positivity), lighter blue (low negativity), dark blue (high negativity; -15).

appear to be sensitive to sentences without the complementizer and in fact a larger effect than the known morphosyntactic agreement violation as shown in difference waveforms in Figure 2-4(row D).

Figure 2-4(row D) shows differences waves for the effect of preferences across all four conditions. The Complementizer, Subject-Verb Agreement, and Gender Reflexive conditions evoke positivity in the P600 time window, and the Lexical Semantic condition evokes the expected negativity within the N400 time window.

These difference waveforms are then plotted spatially on the scalp across time in intervals of 0.2 seconds (Figure 2-5). The warmer colors (e.g., yellow and orange) indicate positive voltages, whereas cooler colors (e.g., green and blue) indicate negative voltages. A P600 response has been evoked when the yellow/orange colors appear between 0.5-1 seconds.

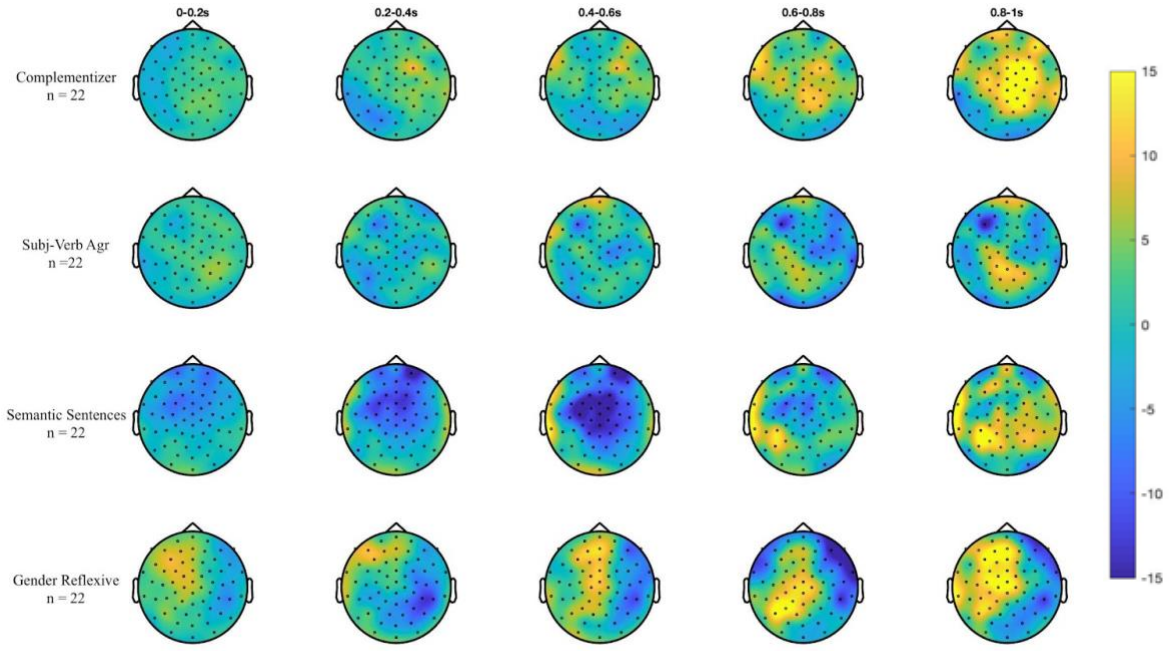


*Figure 2-5: Scalp Topography results in all four conditions over time (0-1 seconds).* Voltage is represented by color; cooler colors are negative values, while warmer colors are positive values. The complementizer condition elicits a stronger P600 response in 0.8-1seconds, while the subject-verb agreement condition shows a weaker P600 response in the same timeframe. The gender reflexive condition elicits a P600 response between 0.6-1.0 seconds. The lexical semantic condition elicits an N400 response between 0.4-0.6 seconds. [Electrodes: Pz, CPz, Cz, CP3, CP1, P1. n=22.]

As shown in Figure 2-5 and the results of the non-parametric cluster analysis, the lexical semantic condition shows a strong centrally distributed N400 effect from 0.4-0.6 seconds. The complementizer condition shows more positivity beginning from 0.6-0.8 seconds, while the response is centrally distributed and the strongest from 0.8-1seconds. In contrast, the subject-verb agreement condition shows positivity posteriorly from 0.8-1 seconds. This difference in scalp topography among the conditions that are expected to evoke a P600 response may suggest that the positivity resulting from the complementizer condition is a syntactic preference violation, whereas the positivity resulting from the subject-verb agreement condition is an outright syntactic violation (Swaab et al., 2011).

The gender reflexive condition shows a P600 effect beginning from 0.6-0.8 seconds with the strongest area from 0.8-1 seconds. The distribution of the gender reflexive is more complicated as it moves from a posterior-central distribution (0.6-0.8 seconds) to a more centrally to left-frontally distribution (0.8-1 seconds). The first timeframe suggests an outright syntactic violation, while the second timeframe suggests a syntactic preference violation.

### *2.3.3 Relationship of the Explicit and Implicit Measures*

#### 2.3.3.1 ERP-Images

The incongruous preferences for the ERP-images (Figure 2-4 row B) visualize the relationship between acceptability and the amplitude of the P600 response. See appendix B for ERP images containing both congruous and incongruous preferences in each condition. If we focus on the relationship between acceptability and amplitude in Figure 2-4 B3, we can see that incongruous preference of the gender reflexive condition appears to have a mismatch from the predicted relationship. We anticipate that lower acceptability is predicted to occur with a larger P600 response, whereas higher acceptability is predicted to occur with a smaller (or no) P600 response. In Figure 2-4 B3, when the trials are rated as 4 (acceptable), the implicit measure frequently shows greater positivity—this is not consistent with a simple proportional relationship[15]. To statistically evaluate the linear relationship between implicit and explicit measures, I ran a linear mixed effect model.

[15] Higher acceptability is predicted to occur with a smaller (or lack of a) P600 response, whereas lower acceptability is predicted to occur with a larger P600 response.

### 2.3.3.2 Results for the Linear Mixed Effects Model

Figure 2-6 illustrates the relationship between acceptability and amplitude for every trial for each participant across all four conditions. No linear relationship is observed for Lexical Semantic violations, which is expected as this condition does not evoke a P600 response.

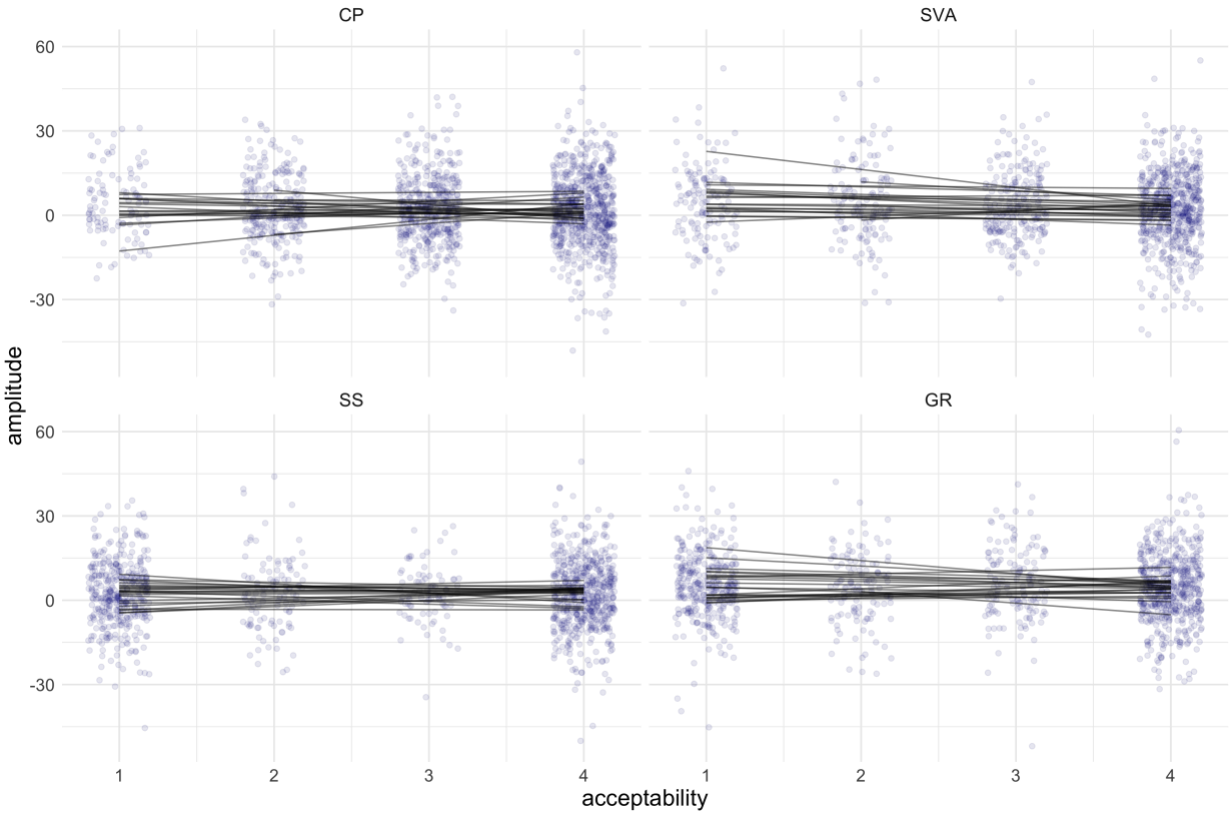


*Figure 2-6: Relationship between Acceptability and Amplitude.* Each panel represents one of the four conditions in the experiment: Complementizer (CP), Gender Reflexive (GR), Lexical Semantic Sentences (SS), and Subject-Verb Agreement (SVA). Each panel shows acceptability judgments across the x-axis and amplitude in microvolts across the y-axis. The blue dots represent the acceptability judgments & amplitude for every trial for each participant. The black lines represent the linear trend for each participant. [Electrodes: Pz, CPz, Cz, CP3, CP1, P1. n=22.]

Across the Complementizer, Subject-Verb Agreement, and Gender Reflexive conditions, there does not appear to be a consistent relationship between amplitude and acceptability across participants. The slopes of the participants trend range from flat, slightly positive, and slightly negative. See appendix C to examine the relationship between amplitude and acceptability for each individual subject.

To assess if the visual relationships in Figure 2-6 are statistically reliable, a linear mixed effects model was fit between single-trial ERP amplitudes and acceptability judgments. Random intercepts and slopes per condition were included. Table 2-12 shows, for each condition, the acceptability judgment trend, and the lower & upper Highest Posterior Density (HPD)[16].

*Table 2-12: Statistical Summary of the Linear Trend per Condition.* The acceptability trend, lower HPD, and upper HPD are slopes across the four conditions in the experiment: Complementizer (CP), Subject-Verb Agreement (SVA), Lexical Semantic (SS), and Gender Reflexive (GR). The model uses data from the P600 time window.

| Condition | Acceptability Trend | Lower HPD | Upper HPD |
|---|---|---|---|
| Complementizer | -0.3835 | -1.077 | 0.3398 |
| Subject-Verb Agreement | -0.8210 | -1.531 | -0.0691 |
| Lexical Semantic | 0.0442 | -0.485 | 0.6331 |
| Gender Reflexive | -0.3344 | -0.936 | 0.2361 |

This analysis suggests that a negative trend (e.g., more acceptable sentences yield less positive ERP amplitudes) is observed for the Subject-Verb Agreement condition, but no reliable trend was observed for any of the other constructions. That is, the estimated linear relationship between acceptability and ERP amplitude was indistinguishable from zero for Complementizers, Gender Reflexives, and Lexical Semantic sentences.

## 2.4 Discussion

Through collecting and analyzing the online and offline data in a single trial, I tested whether gradient acceptability judgments to a target construction show proportional gradient implicit neural responses to the P600. I had predicted that this experimental approach would better characterize what the relationship between these online and offline measures are. The results suggest that there is not a reliable negative correlation between ERP amplitude and acceptability judgments across all conditions, only the SVA condition. In section 2.4.1, I discuss how the implicit and explicit results compared to the predictions (2.2.4). Section 2.4.2 discusses

[16] The upper and lower HPD are the boundaries for statistical reliability. For a proportional relationship, the slopes of the acceptability trend, lower HPD, and upper HPD would all be negative, or they would all be positive.

analyzing the relationship between the implicit and explicit measures. Section 2.4.3 will discuss what the online and offline results suggest for the complementizer condition and the status of it in grammar of the participants. Finally, section 2.4.4 considers if the RSVP reading protocol affects the reliability of the implicit neural signal.

### *2.4.1 Implicit and Explicit Results*

The acceptability judgment results mostly aligned with the predictions in section 2.2.4, except for the subject-verb agreement condition. The incongruous preference in the subject-verb agreement[17] acceptability judgments exhibited gradience, which differed from the norming study and the current study's predictions. This was due to procedural differences, as participants in the EEG study were able only to see the sentence once and rate the acceptability of the sentence within 4000 ms (Figure 2-1). The perception in the incongruous subject-verb agreement condition is noisy—perhaps due to the ability to notice and hold the violation in memory. These EEG experiment acceptability judgment results replicated the reported behavioral results in Tanner (2019) using a 4-point Likert-scale, rather than binary choice (e.g., "good" and "bad"). In fact, Tanner found that only the accuracy within the incongruous subject-verb agreement condition was significantly lower than all other conditions in both RSVP and self-paced reading protocols. "These data indicated that participants had a more difficult time responding accurately to errors in the ungrammatical lexical verb condition than all other conditions, and that task was not a reliable determinant of judgment accuracy" (Tanner, 2019, pp. 219–220). These data from my own experiment provide further confirmation of Tanner's (2019) results using a different

[17] Tanner (2019) used the label subject-(lexical) verb agreement to differentiate from another condition using auxiliary verb disagreements.

type of acceptability judgment scale, the same lexical subject verb experimental stimuli, and RSVP procedure.

Overall, the predicted ERP components aligned visually, such that the N400 was elicited to the lexical semantic condition and the P600 was elicited for the CP, SVA, and GR conditions. However, the predicted ERP effects were nonsignificant for CP, SVA, and GR conditions, but they were significant for the lexical semantic condition. It is possible that this EEG experiment needed more than 22 included participants to account for the noise in the ERP data and the percentage of the excluded trials (Table 2-11)[18]. It is likely that the non-significance in the online measures of the conditions predicted to elicit a P600 response obscured the relationship between acceptability judgments and the strength of the ERP amplitude.

### *2.4.2 The Relationship between the Implicit and Explicit Measure*

To visualize the relationship between the implicit and explicit measures, I used the ERP-image (Delorme et al., 2015; Sassenhagen et al., 2014) that includes four dimensions within a single image: time course, trials, amplitude, and acceptability judgements. The advantage of using the ERP-image is that gradience in the implicit and explicit measures is somewhat easy to identify within the ERP images. The implicit measure shows a wider range of color representing the amplitude, while the explicit measure is represented by the black line for all the trials. The length of the line indicates how many trials answered that value. Overall, the ERP-images in Figure 2-4 (row B) show strong P600 responses across all four Likert-Scale points, which suggests a lack of a proportional relationship between ERP amplitude and acceptability. This is exactly what we see with the results of the linear mixed effect model in Table 2-12.

[18] Höller (2021) discussed that EEG studies historically use 20 participants, although studies now range from a small to large number of participants. Höller suggests conducting power analyses to determine the number of participants needed for EEG studies.

A reliable negative trend (e.g., more acceptable sentences yield less positive ERP amplitudes) is only observed for the Subject-Verb Agreement condition. The estimated linear relationship between acceptability and ERP amplitude was indistinguishable from zero for the other three conditions. As the online measures were not statistically reliable in the P600 conditions, it could be affecting the relationship—perhaps that the relative strength of the N400 and P600 response in these conditions is affecting the ERP response (see Chapter 3).

The lack of relationship between the online and offline measure in the lexical semantic condition was expected, as the condition evoked a statistically reliable N400 response, but the ERP data used in the linear mixed effect model was within the P600 time window. This leaves open the possibility that, in the lexical semantic condition, the relationship between ERP amplitude of the N400 and acceptability judgments could be proportional if tested with data from the N400 time window.

### *2.4.3 Gradience in Acceptability and the Variable Complementizer Condition*

The variable complementizer condition shows gradience in acceptability (Figure 2-3) as the by-subject averages show acceptability judgment ratings across all four points of the Likert-Scale in the incongruous complementizer condition. But more importantly, the grand average acceptability rating between the congruous and incongruous preferences are not statistically different. Participants individual differences are examined in Appendix A, and this suggests that subjects differ in whether they rated the two preferences in the complementizer construction as categorical, gradient, or the same[19]. These offline acceptability judgment results do not provide a clear picture about if the construction is part of the grammar.

[19] Participants R0649, R0650, and R0721 show gradience in acceptability for the congruous and incongruous complementizer conditions, while only one participant (R0684) showed categorical judgments. Participants R0670,

The online ERP amplitude results also do not provide a clear-cut explanation of whether the variable complementizer construction is part of the grammar as there was only a trend toward a P600 response, not a statistically significant difference. The incongruous ERP amplitude in the complementizer condition is similar in magnitude to the incongruous subject-verb agreement and gender reflexive conditions (Figure 2-4), which would provide support for the strong violation, the P600 amplitude to sentences without the complementizer would be of a similar magnitude of a known ungrammatical violation (e.g., subject-verb disagreement). But the central positive distribution of the incongruous complementizer condition in Figure 2-5 suggests a syntactic preference violation (Swaab et al., 2011). Taking the online and offline measures together, this study provides inconclusive evidence on whether the variable complementizer condition is part of the participants grammar.

### *2.4.4 Methodological Consideration*

The online ERP effects may have been affected by the RSVP reading protocol in the experimental procedure, as participants may have been less engaged with accurately reading sentences that appeared automatically. Bulkes et al. (2020) suggested that self-paced reading (SPR) may be more engaging than RSVP reading protocols, as participants are able to dictate the pace of the stimuli presentation. Tanner (2019) tested the reliability of using the RSVP and SPR reading protocols and found that the online ERP effects were reliable in sentences containing a violation using both reading protocols, as the grand average waveform analyses did not yield statistically significant differences. Since the conditions eliciting the P600 response in this

R0695, R0472, R0493 rated the congruous and incongruous complementizer conditions as acceptable, while participants R0719 R0699 rated both preferences as unacceptable.

experiment were not significant, the change to the SPR protocol may alter the online ERP effects, and thus provide clearer insight on whether the relationship between acceptability judgments and P600 amplitude is proportional.

### 2.5 Summary of the Study

This study tests whether gradient acceptability judgments to a target construction show proportional gradient implicit neural responses to the P600. Participants read and rated the acceptability of each sentence across four conditions: complementizer, subject-verb agreement, gender reflexive, and lexical semantic. A linear mixed effects model was fit with amplitude as a dependent measure and acceptability, condition, and their interaction as a fixed effect. The model included subjects as a random intercept and random slopes for their acceptability, condition, and their interaction. Results indicate that the CP, SVA, and GR conditions show a negative acceptability judgment trend. However, only the SVA condition shows statistically reliability in a negative relationship between amplitude and acceptability. Overall, the results do not support a proportional relationship between implicit neural and explicit acceptability judgment measures.

## Chapter 3 Examining Individual Differences in Implicit Neural Measures

This analysis examines whether there is evidence of individual differences in the ERP data across the conditions that may contribute to gradience of the implicit neural signal. A productive avenue for inquiry is on the relative strength of two distinct ERP response, the P600 and N400 (see Table 2-1 in Chapter 2). Prior work suggests that individuals may differ, in a gradient way, in exactly this balance (Tanner, 2019; Tanner & Van Hell, 2014; Zeitlin, 2020). This chapter explores the individual differences present in the ERP data from Chapter 2 to better understand the extent of gradience within the implicit neural signal.

### 3.1 Motivation and Research Question

Variation in the relative prominence of the N400 and P600 responses in condition that traditionally evoke a P600 response could lead to gradience in the P600 response, which in turn, may affect the relationship between the online ERP and offline acceptability judgment measures. Prior research has used response dominance index to examine the relative prominence of the N400 and P600 across participants for auxiliary and lexical verb agreement violations (Tanner, 2019; Tanner & Van Hell, 2014). The Response Dominance Index (RDI) is used to determine the relative prominence of the N400 and P600, and the RDI is calculated in two steps (Equation 1). First, the magnitude of the averaged ERP data is calculated in the P600 time window (500-800 ms) and the N400 time window (300-500 ms), which is subtracted from the P600 magnitude.

*Equation 1: Response Dominance Index (RDI) (Tanner, 2019)*[20]

$$RDI = \frac{\left(P600_{Ungram} - P600_{Gram}\right) - \left(N400_{Gram} - N400_{Ungram}\right)}{\sqrt{2}}$$

Second, this singular magnitude value is transformed by dividing the magnitude by the square root of 2, revealing the RDI values. Negative values indicate more negativity across these two time-windows, positive values indicate more positivity across these two time-windows, and near zero indicates roughly equal distribution of negative and positive values within these time windows. Previous research indicated that individual RDIs scores tended to show similar polarity responses for both auxiliary and lexical verb agreement violations (Tanner, 2019; Tanner & Van Hell, 2014) meaning the participants were processing these agreement violations similarly within the N400 and P600 time windows.

Processing the same linguistic stimulus across the time course may yield different processing strategies that are evident in the specific ERP effects. N400 effects are likely to represent retrieving and integrating information from long-term semantic memory (Lau et al., 2008; Tanner, 2019; Zeitlin, 2020), and P600 effects are likely monitoring the input and processing repairs or reanalysis (Osterhout & Nicol, 1999; Tanner, 2019; Zeitlin, 2020). Tanner states that "as the dominance of one set of processes increases, the dominance of the other decreases" (2019, p. 227). These processing strategies within the N400 and P600 time windows are likely to affect the RDIs at the group and individual level[21].

The ERP experiment in Chapter 2 has shown gradience in acceptability judgments in the complementizer condition and the subject-verb agreement condition. However, it is unclear if

[20] The terms ungrammatical and grammatical should be thought of as predicted grammaticality within the dissertation. Within this dissertation, incongruous aligns with ungrammatical and congruous aligns with grammatical.

[21] Prior work has shown that individual differences in cognitive function (e.g., working memory (Kim et al., 2018)) also modulates the N400 and P600 components, which could affect RDIs at the group and individual level.

participants implicitly process these conditions similarly in that participants evoke the expected neural signal, because the main effects of preference are not statistically reliable according to the non-parametric cluster analysis (see Chapter 2). To assess if participants are implicitly processing these conditions similarly, I examine the violations that typically evoke a P600 response (e.g., Subject Verb Agreement or Gender Reflexive violations), do participants individually show positive-going RDIs in conditions that traditionally evoke a P600 response? This conceptual replication would show congruency with the literature and neural signal of the participants. A crucial next question is whether individual participants show a stable RDI across other violations that traditionally evoke a P600 response? A stable RDI would be defined as maintaining the same RDI (e.g., positive-going, negative-going, or biphasic) across the other violations evoking a P600 response). To provide evidence that the Complementizer condition is processed like constructions that traditionally evoke a P600 response, we would expect a participant's RDI to remain stable across the Gender Reflexive and Subject-Verb Agreement conditions (e.g., positive-going RDI in both, negative-going RDI in both, or biphasic RDI in both).

## 3.2 Methods

The neural data used came from the study reported in Chapter 2. For this study, the methods section will focus on the predictions (3.2.1) and the data analysis procedure (3.2.2).

### *3.2.1 Predictions*

Table 3-1 provides a summary of the predicted RDIs across each condition based on the ERP results in Chapter 2.

*Table 3-1: Predicted Response Dominance Index in Each Condition based on Chapter 2 ERP Results.* (See Section 3.4.1 for a discussion on literature predictions).

| Condition | ERP Component Evoked in Chapter 2 | Predicted Response Dominance Index |
|---|---|---|
| Gender Reflexive | P600 | Positive |
| Subject-Verb Agreement | P600 | Positive |
| Complementizer | P600 | Positive |
| Lexical Semantic | N400 | Negative |

The Gender Reflexive, Subject-Verb Agreement, and Complementizer conditions evoked a P600 response in Chapter 2, and the Lexical Semantic condition evoked an N400 response. For the conceptual replication, the predicted RDI will align with the polarity of the ERP component from the neurolinguistic literature (see section 2.1.1). The three conditions that evoked a P600 response are predicted to show positive RDIs in participants. The lexical semantic condition is predicted to show a negative RDI in participants due to the statistical reliability found in Chapter 2. For the novel extension, I predict that these positive RDIs will remain stable across the conditions that evoke a P600 response.

### *3.2.2 Data Analysis Procedure*

I follow the analyses set forth by Tanner (2019) and Tanner and van Hell (2014) to test the magnitude effect and the RDI within each condition[22]. The averaged ERP data from the N400 (300-500 ms) & P600 (500-800 ms) time windows in the centro-parietal region of interest were selected. The centro-parietal region of interest was defined by the following electrodes: C3, Cz, C4, CP1, CP2, P3, Pz, P4. This time window data was used in measurements for (1) magnitude of the P600 and N400 effects and (2) the RDI (Equation 1). The magnitude of the P600 effect is calculated by subtracting the grammatical from the ungrammatical stimuli in the P600 time window. The magnitude of the N400 effect is calculated by subtracting the ungrammatical from

[22] Tanner (2019) reported the results for the lexical subject-verb agreement and auxiliary subject-verb agreement conditions. The gender reflexive and lexical semantic conditions were fillers and these RDIs were not reported.

the grammatical stimuli in the N400 time window. These N400 and P600 magnitude effects were passed to a correlation analysis within each condition (results in Table 3-2) to determine the relationship between these magnitudes. The N400 and P600 magnitude effects are used in the numerator of Equation 1 for the RDI.

## 3.3 Results

### *3.3.1 Magnitude Effects*

Figure 3-1 shows the individual magnitudes for the N400 and P600 effects across all four conditions. The quadrants of the graph indicate the primary effect each participant maintained within that condition. Data points in the upper left quadrant represent a sustained negativity through the 300-800 ms time window, which indicates that there is no P600 effect. Within the same time window, the data points in bottom right quadrant represent an early sustained positivity without an N400 effect. Data points located in the upper right quadrant indicate a biphasic response (i.e., equal N400 and P600 response). It is not expected that data points would fall within the lower left quadrant as this would indicate a negative P600 and N400 effect amplitude. Data points in this quadrant are likely noise from the ERP data.

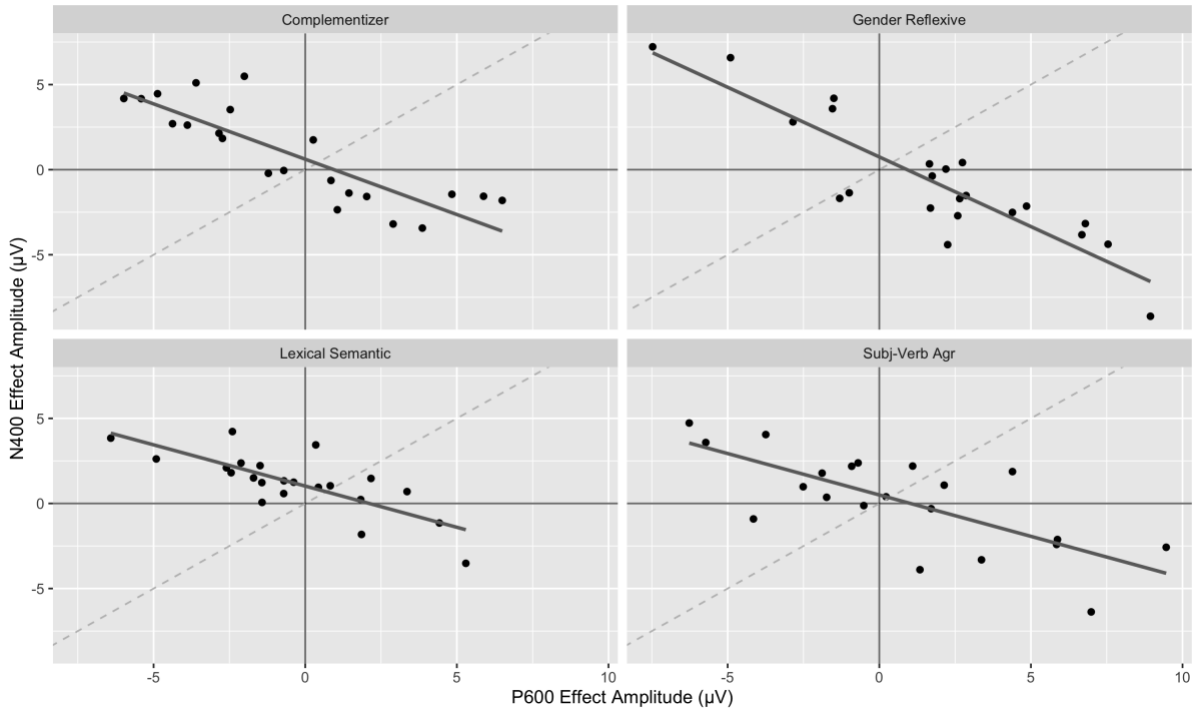


*Figure 3-1: Individual Effect Magnitude in all Conditions.* The relationship between the P600 effect amplitude (ungrammatical minus grammatical; x-axis) and N400 effect amplitude (grammatical minus ungrammatical; y-axis) is shown for each participant across all four conditions. The solid black line shows the least squares best fit line. The dotted grey line represents the equal effects line where the P600 and N400 effect amplitudes are equal. Individuals above/to the left of the equal effects line demonstrate a negativity-dominant brain response, while individuals below/to the right of the equal effects line demonstrate a positivity-dominant brain response. [Electrodes: C3, Cz, C4, CP1, CP2, P3, Pz, P4; n=22].

The Complementizer condition and the Subject-Verb Agreement condition shows an equal number of participants in the upper left and lower right quadrants. About half of the participants sustained negativity throughout the time window and the others sustained an early positivity. This suggests that, across participants, there was relative balance between evoked N400s and evoked P600s for these conditions. This result differs from Tanner (2019) as the SVA condition had a majority of participants within the lower right quadrant—suggesting a sustained early positivity. The Gender Reflexive condition shows many of the participants in the lower right quadrant with a sustained early positivity. In contrast, the Lexical Semantic condition, which evokes an N400 response, shows most participants clustered in the upper left quadrant indicating sustained negativity.

The N400 and P600 magnitude effects in each condition were passed to a correlation analysis and the results are shown in Table 3-2.

*Table 3-2: Correlation Results of N400 and P600 Magnitude Effects within each Condition.*

| Condition | Correlation | Significance Value |
|---|---|---|
| Complementizer | -0.8497504 | $p < 0.0001$ |
| Subject-Verb Agreement | -0.7378217 | $p < 0.0001$ |
| Gender Reflexive | -0.8863292 | $p < 0.0001$ |
| Lexical Semantic | -0.7659055 | $p < 0.0001$ |

The correlation analyses returned a negative value and a significant p-value (< 0.0001) for each condition. This means across all four conditions, if individuals showed a large P600 effect, then they tended to show little negativity and vice-versa. This aligns to the visual representation in Figure 3-1, in which most of data points fell within a sustained negativity (upper left quadrant) or sustained early positivity (lower right quadrant). Very few data points maintained a biphasic response (upper right quadrant). These results link back to how participants may be processing the stimuli (e.g., semantic integration or syntactic reanalysis) and suggest that participants are doing one type of processing for these sentences—either the N400 or P600, which indicates that there should be minimal biphasic participants.

### *3.3.2 Individual Response Dominance Index*

Figure 3-2 shows the stability of the RDI values across all four conditions for each participant. RDI values that are greater than 1 show a dominance of positivity and RDI values that are less than -1 show a dominance of negativity. If participants show biphasic response (roughly equal-sized N400 and P600 effects), their RDI values will fall between -1 and 1. The Gender Reflexive condition is compared to the Subject-Verb Agreement condition to see if RDIs

remained stable across conditions that typically evoke a P600 response in the literature. The complementizer condition evoked a P600 response in Chapter 2. The Lexical Semantic condition is not used to assess if the participants RDIs remain stable, as this condition evokes an N400 response and not a P600 response.

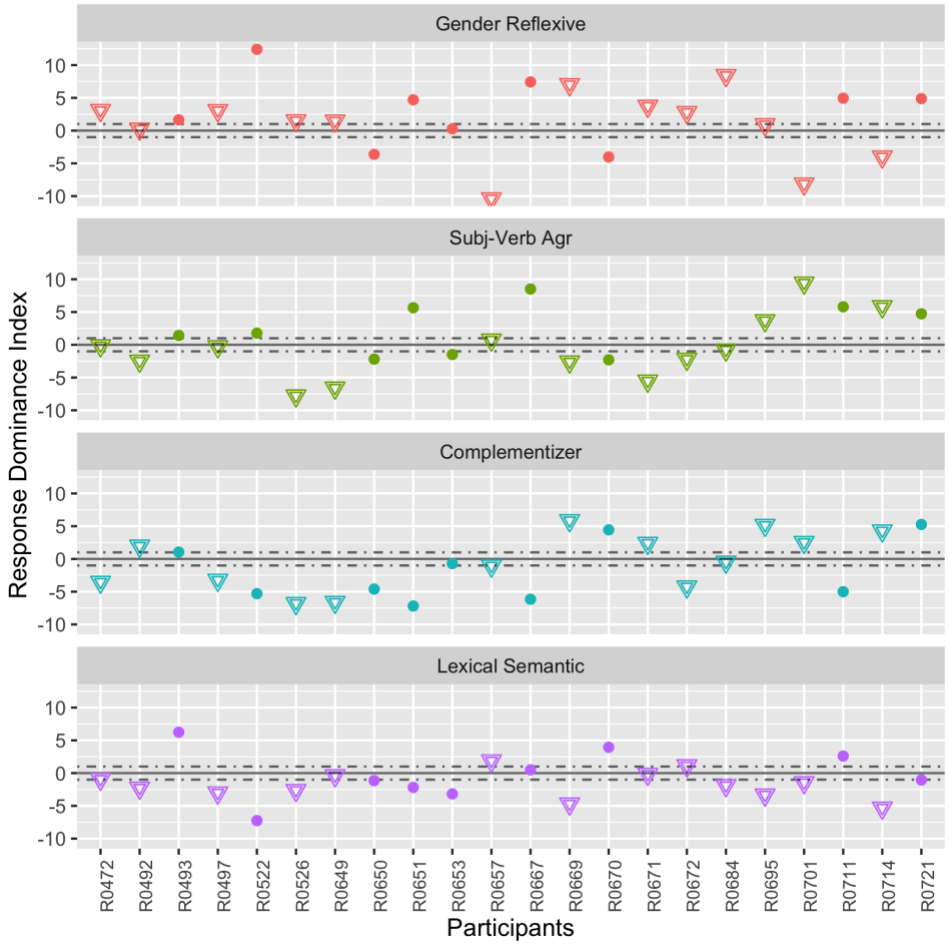


*Figure 3-2: Stability of Response Dominance Index in all Conditions across Participants.* Participants are located across the x-axis and the Response Dominance Index is located on the y-axis. The dotted lines are located between -1 and 1, which indicates a biphasic RDI. Values above 1 indicate a positive-going RDI and values below -1 indicate a negative-going RDI. Dots represent a stable response from the Gender Reflexive condition as compared to the Subject-Verb Agreement condition. The triangles represent a change in dominance response from the Gender Reflexive condition as compared to the Subject-Verb Agreement condition. [n=22].

In the Gender Reflexive Condition, most participants demonstrate a positive-going or biphasic response (14 and 3 participants respectively), and only 5 participants are showing a negative-going response. The individual RDIs to the subject verb agreement violations shows fewer positive-going responses than in gender reflexive violations, which impacts the stability of

the RDIs across the individual participants. The SVA condition is split between the positive and negative RDIs at 9 participants. Changes in the RDIs from the Gender Reflexive violations to the Subject-Verb Agreement violations are represented by the triangles. A total of 13 participants exhibited different RDI responses in these two conditions, while 9 participants maintained their RDI responses. This evidence does not support a stable response in the two conditions that traditionally elicit a P600 response for this data.

Using the visualization from Figure 3-2, I counted the participants by their dominant response in each condition. Table 3-3 reports the number of participants that fall within the negative-dominant, positive-dominant, and biphasic groups across the four conditions.

*Table 3-3: Response Dominance Index Groupings.* Participants were grouped in each condition by their Response Dominance Index. Near zero participants were decided by if the RDI fell between -1 to 1.

| Condition | Number of Negative-Dominant Participants (Negative Values) | Number of Positive-Dominant Participants (Positive Values) | Number of Biphasic Participants (Values between -1 to 1) |
|---|---|---|---|
| Complementizer | 11 | 9 | 2 |
| Subject-Verb Agreement | 9 | 9 | 4 |
| Gender Reflexive | 5 | 14 | 3 |
| Lexical Semantic | 13 | 5 | 4 |

Table 3-3 shows the number of participants varies. This indicates unstable RDIs among the constructions that typically evoke a P600 response, as the number of participants should be similar for these constructions. The complementizer condition and the subject-verb agreement condition show equal sized groups for negative or positive dominances. The conditions that show a majority in their dominance is the Gender Reflexive and the Lexical Semantic conditions. These conditions show the majority in the predicted ERP component, such that there are more positive dominant participants for the Gender Reflexive condition (evoking a P600) and there are more negative dominant participants for the Lexical Semantic condition (evoking an N400).

As only 9 participants showed stability in their RDIs from the Gender Reflexive to the Subject-Verb Agreement, we can see how many individuals maintained their RDIs across the three conditions[23] by also comparing to the Complementizer condition. Figure 3-3 visualizes the RDIs from all conditions by individual subject and shows stability the P600 constructions if they are located within one area of the graph. For example, if the RDIs were positive-going, the dots would be located above the dotted line at y=1. If the RDIs were Negative-going, the dots would be located below the dotted line at y= -1. And biphasic RDIs would lie between the dotted lines (between 1 and -1) on the graph.

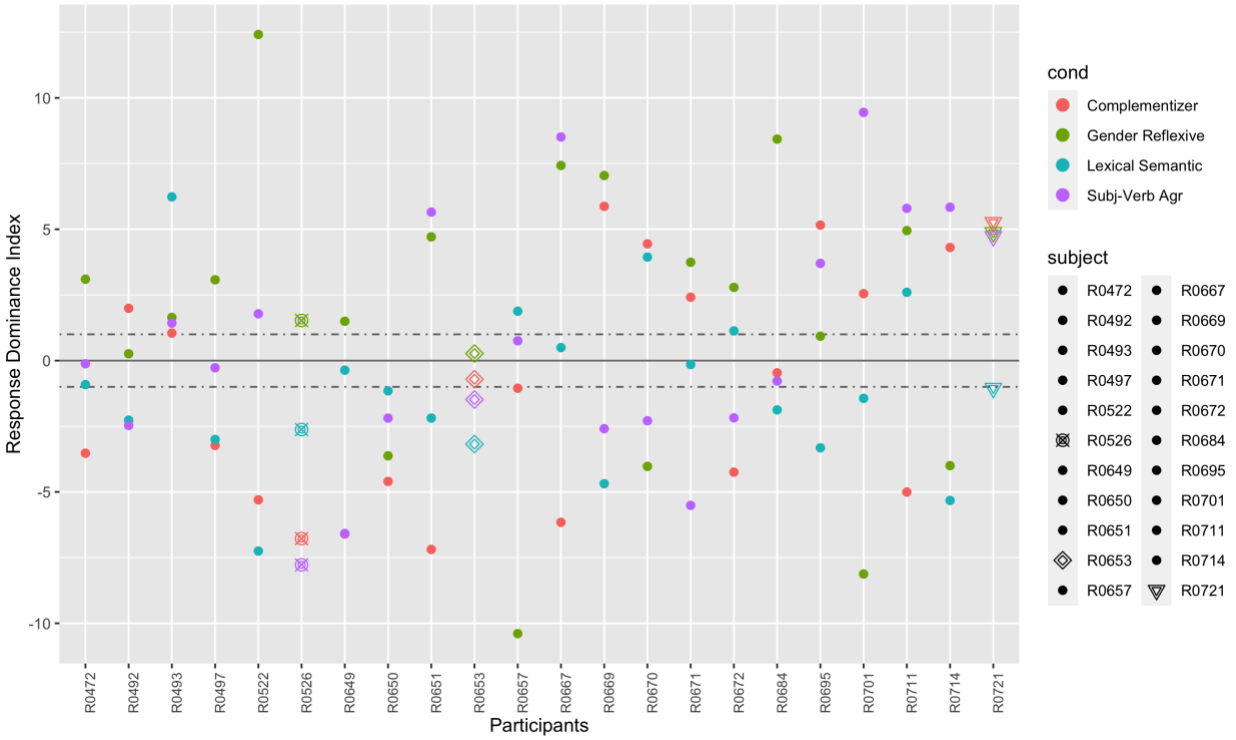


*Figure 3-3: Individual Response Dominance Indexes Across Subject and Conditions*. Participants are located across the x-axis and the Response Dominance Index is located on the y-axis. The dotted lines are located between -1 and 1, which indicates a biphasic RDI. Values above 1 indicate a positive-going RDI and values below -1 indicate a negative-going RDI. Three subjects (R0526, R0653, and R0721) are shown in different shapes to highlight the different kinds of RDIs across these conditions. [n=22].

[23] Note that the color of the conditions in Figure 3-3 are different from the condition colors in Figure 3-2.

The RDIs across the three conditions that evoked a P600 response in Chapter 2 is unstable. Two participants showed stability across the three conditions. Participant R0721 shows a positive-going RDI across the 3 conditions, which is in line with the predictions in section 3.2.1, while Participant R0650 shows a negative-going RDI across all three conditions. Two participants (R0493 and R0695) are close to having stable responses across the three conditions, but one condition borders on the biphasic group cutoff. The rest of the participants show unstable responses across the Gender Reflexive, Subject-Verb Agreement, and Complementizer conditions. One construction is not responsible for the instability across participants. For example, R0526 shows a negative-going RDI for the Complementizer and Subject-Verb Agreement conditions, but a positive-going RDI for the Gender Reflexive condition. Participant R0653 shows a biphasic response for the Gender Reflexive and Complementizer conditions, but a negative-going RDI for the Subject-Verb Agreement condition.

### *3.3.3 ERP Waveforms Based on RDI Groupings*

Aside from the lexical semantic condition, the ERP measures from Chapter 2 were not reliable; however, grouping by RDI may alter the gradience present within the ERP responses[24]. Figure 3-4 explores this possibility by taking the RDI groupings from the previous section (3.3.2) and computes the grand average ERP waveforms by dominance grouping[25]. Row A is discussed in section in Chapter 2.

[24] This could then yield significant effects for CP, SVA, and GR when grouping by RDI. In turn, this could then affect the relationship between online and offline measures and perhaps clarify if ERP amplitude and acceptability may correlate reliably.

[25] These groupings are not adequately powered to make reliable conclusions or broad generalizations about the effect of RDIs on the ERP waveforms. However, these graphs lend us a useful visual to hypothesize about how RDI groups may inform future experiments that test the RDI of individuals.

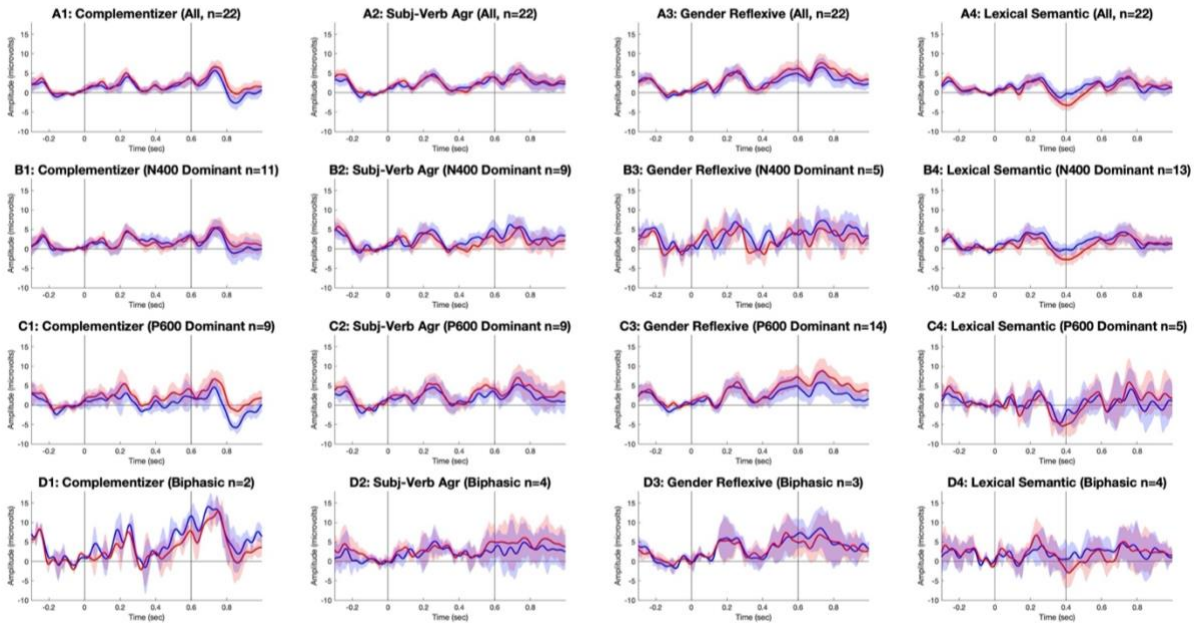


*Figure 3-4: Grand Average Waveform ERP Results by Dominance Grouping.* Time (seconds) is represented across the x-axis and amplitude (microvolts; positivity plotted up) is listed across the y-axis. Grey vertical lines are displayed at 0 seconds (stimulus onset) and again at either 0.4 seconds (predicted peak N400 response) or 0.6 seconds (predicted P600 response). The congruous preferences are shown in blue (e.g., predicted grammatical), and the incongruous preferences are shown in red (e.g., predicted ungrammatical). Row A shows the grand average waveform for all participants across all four conditions, Row B the negativity-dominant group, Row C the positivity-dominant group, and Row D the biphasic group. [Electrodes: C3, Cz, C4, CP1, CP2, P3, Pz, P4].

Figure 3-4 (Row B) is grouped by the negative-going RDIs, and largely the incongruous preference (red) is below the congruous preference (blue) within the 300-500 ms time window. The clearest N400 response is shown in the Lexical Semantic condition, which is known to evoke an N400 response. The Subject-Verb Agreement condition also shows a sustained negativity throughout the time course. Figure 3-4 (Row C) shows the incongruous preference (red) above the congruous preference (blue). The Complementizer and Gender Reflexive conditions show a sustained early positivity in these conditions, with the P600 response starting at 300 ms. The biphasic group in Figure 3-4 (Row D) is not interpretable from the small number of participants. We would expect to see patterns of negativity from 300-500 ms and positivity from 500-800 ms. Overall, grouping the data by RDI groups shows clearer N400 responses in the negative-going groupings, P600 responses in positive-going, and small biphasic responses.

However, the standard errors for the congruous and incongruous preferences significantly overlap throughout the time course, following the results in Chapter 2.

### 3.4 Discussion

This study investigated if participants individually showed positive-going RDIs in violations that traditionally evoke a P600 response. The novel extension takes this further to explore if individual participants exhibit stability in their RDIs across other violations that traditionally evoke a P600 response. At first glance, there is confirmation of a conceptual replication, as the Gender Reflexive violations show a majority of positive-going RDIs (n=14). The Subject-Verb Agreement violations show an equal number of positive-going and negative-going RDIs (n=9). This conceptual replication, while not in line with my predictions in section 3.2.1, shows congruency with the literature and neural signal of the participants (as discussed in section 3.4.1 below).

While the results indicate a conceptual replication of the literature, the novel extension in determining whether the individual RDIs are stable as compared to conditions typically eliciting a P600 response failed. Over half of the participants (n=13) showed a change in their RDI indicating instability in their RDIs across the gender reflexive and subject-verb agreement violations. This study does not support a stable response in two conditions that traditionally elicit a P600 response. The result follows recent work from Zeitlin (2020), who found that an individual's RDI to one syntactic violation did not predict stability to any other syntactic condition. This difference in RDIs among individuals could be due to two factors: (1) other ERP components within the constructions and (2) the ranking of language processing strategies following what violations elicit an N400 or P600 component (refer to Table 2-1).

### *3.4.1 Conditions and Multiple ERP Components*

One issue that may affect these RDI results is the fact that the violations that traditionally evoke a P600 response are susceptible to evoking other ERP responses. Table 3-4 shows the possible ERP components that may also be elicited for the conditions in this experiment.

*Table 3-4: Possible Language-Specific ERP Responses Affecting the Conditions that Typically Exhibit P600 Responses*. The Gender Reflexive and Subject-Verb Agreement conditions may also exhibit an AN or LAN response according to the literature (Luck, 2014; Swaab et al., 2011). The Complementizer condition hasn't been tested in other ERP studies, so it is unclear if these ERP components may be elicited in this condition.

| Condition | Violations in the Condition | Results from Chapter 2 | Other Possible ERP Components |
|---|---|---|---|
| Gender Reflexive | Gender agreement violation | P600 | LAN or AN |
| Subject-Verb Agreement | Number Agreement violation | P600 | LAN or AN |
| Complementizer | Perhaps a phrase structure violation or a dispreferred syntactic structure violation | P600 | Early Left Anterior Negativity[26], LAN, or AN |

LAN/AN are negative-going waves that occur 300-500ms post stimulus, which is the exact timeframe as the N400 magnitude effects. Violations in agreement, case, phrase structure, island constraints may elicit a LAN/AN component, in addition to garden-path sentences and long-distance dependencies (Luck, 2014; Swaab et al., 2011). In this study, the Gender Reflexive and Subject-Verb Agreement conditions may also evoke a Left Anterior Negativity (LAN) or an Anterior Negativity (AN). Tanner (2019) found that lexical subject-verb agreement verbs elicited an enhanced N400 relative to auxiliary subject-verb agreement, which occurs in the same time window as the LAN/AN components. This is likely why the Subject-Verb Agreement condition was split evenly between a negative-going RDI and a positive-going RDI. This split between

[26] Early Left Anterior Negativity (ELAN) is an ERP component that occurs between 100 ms to 250 ms post stimulus to phrase structure violations or word category violations (Swaab et al., 2011). An ELAN may appear in constructions eliciting a P600, but the P600 may appear without an ELAN.

negative-going and positive-going RDIs is likely due to various processing strategies (e.g., integration or reanalysis) within the N400 and P600 time windows.

### *3.4.2 Processing Strategies*

Given that the number of participants in the study is small, examining the data from the RDI groupings is limited. Any widespread generalizations would need to wait for more participants. However, the participants seem to employ different language processing strategies in conditions that typically report P600 effects. The RDI groupings in the Subject-Verb Agreement condition suggest a split in processing strategies among participants. Some participants may have prioritized integrating information (N400 effect) given the negative RDIs, as opposed to reanalyzing the stimulus (P600 effect) given the positive RDIs. This explanation may also explain why the explicit acceptability judgments in Chapter 2 were higher than anticipated. Participants who reanalyzed the construction may have rated the acceptability of the stimulus higher than a participant who only attempted to semantically integrate the violation without syntactic reanalysis.

The instability shown across the gender reflexive and subject-verb agreement violations may diminish if the same types of processing strategies used for each violation are prioritized or ranked similarly. This leads to an open question of why might one processing strategy be prioritized versus another in conditions that typically report P600 effects? One possible direction would be to design another ERP study that focuses on integration and reanalysis strategies in stimuli typically evoking P600 effects. If the processing strategies could be manipulated to focus on either the integration or reanalysis strategy, then we could assess if use of the same processing strategy in violations that traditionally elicit a P600 response would show a stable RDI across individual participants.

### 3.5 Summary of the Study

This study examines two key research questions: (1) do participants individually show positive-going response dominance indexes in violations that traditionally evoke a P600 response and (2) do individual participants show a stable response dominance index across other violations that traditionally evoke a P600 response? ERP data from the N400 and P600 time frames were used to calculate the RDI for each participant in all four conditions. The results indicate that RDI is not stable across the conditions to elicit the predicted P600 response. This suggests that participants employ various cognitive processing strategies (e.g., semantic integration and syntactic reanalysis) within a single condition that is predicted to evoke a P600 response. The usage of different cognitive strategies within and across different conditions that typically evokes a P600 response likely explains the individual differences in RDI among participants and contributes to the gradience in the online measure from Chapter 2.

## Chapter 4 Exploring Gradience in Online Reading Time Measures and Offline Acceptability Measures Using Self-Paced Reading (SPR)

This experiment extends the combination of online (ERP) and offline (judgment) data to a methodology suitable for use with internet-based data collection. The experiment used a self-paced reading protocol to explore the link between offline acceptability judgments and online reading times (RT) in a task targeting sentence acceptability. Specifically: Do gradient acceptability judgments to a target condition show proportional gradience in online RTs? The two measures are compared in two linguistic conditions: i) sentences argued to exhibit gradient acceptability due to variable complementizer realization (Sprouse et al., 2013; Martin, 2001) and ii) sentences known to be ungrammatical in subject-verb agreement (Tanner, 2018 & 2019). In using identical stimuli to Chapter 2, the experimental methodology is expanded to include online RTs to determine how this online measure relate to known conditions exhibiting gradient acceptability.

### 4.1 Motivation for an SPR Study via the Internet

The COVID19 pandemic affected my ability to complete in-person data collection and thus warranted the need to collect data via the internet. The primary issue with the methodology set forth in Chapter 2 was that there would not be an online measure to collect data if the Rapid Serial Visual Presentation (RSVP) reading protocol remained. Alternative reading protocols are available, and different reading protocols have been tested against each other in the literature. Tanner (2019) conducted two ERP studies using two different protocols, RSVP and Self-Paced

Reading within that study, to investigate whether individual differences in the neural signature were a result of various reading protocols used among visual ERP studies.

Research has suggested that self-paced reading may be more engaging than RSVP protocols (Bulkes et al., 2020), as self-paced reading allows participants to dictate the pace of the stimuli presentation. In self-paced reading protocols, the RT to each word is recorded in milliseconds as the online measure. For syntactic conditions, RTs slow one position past the violation in ungrammatical stimuli compared to grammatical stimuli (Tanner, 2019; Vafaee et al., 2017; Van Berkum et al., 2005; Vincenzi et al., 2003), and this slow down indicates higher processing demands on participants.

Tanner (2019) found reliable online ERP effects in sentences containing a violation using both RSVP and SPR reading protocols. The grand average waveform analyses for each type of reading protocol were not statistically significant differences. With respect to RTs, Tanner reports slower RTs for the ungrammatical Subject-Verb agreement condition as compared to the grammatical Subject-Verb agreement condition. These slower RTs occurred at target word+1, but there was not a significant difference at the target word or target word +2 specifically.

Given this evidence, I can adapt my experimental methodology to an online platform by using the self-paced reading protocol and by using the same conditions from Chapter 2. I expect the results from this experiment and the EEG experiment in Chapter 2 to be similar. Incongruous preferences that evoke a P600 response are also expected to show an increase RTs immediately after the target word (+1 relative to the target word).

While the amplitude and RT measures are predicted to behave similarly (e.g., a P600 response and slower RTs), it is not known whether gradient acceptability judgments to a target condition would show proportional gradience in online RTs. This study tackles this question by

minimally altering the experimental methodology to include the self-paced reading protocol in an online study format.

## 4.2 Methods

### *4.2.1 Participants*

A total of 80 participants were recruited on Prolific and participants were compensated at the rate of $9.50 per hour for their participation. Participants were recruited using the following criteria: English as one of their native languages ("First language"), located in the United States, and then prolific equally distributed among genders. Participants had to take the experiment using a desktop/laptop computer.

### *4.2.2 Self-Paced Reading and Acceptability Judgment Procedure*

The experiment consisted of three tasks: a self-paced reading practice block, a main self-paced reading experiment, and a sociolinguistic demographic survey. The experiment was designed to be completed in 45 minutes, but participants were allowed 115 minutes to complete all three tasks. On average, participants took 23 minutes to complete the experiment.

Before the practice and experimental blocks began, participants read an instructions page explaining the mechanics of the experiment and how to rate the acceptability of sentences: "You'll be reading sentences at your own pace. Once you've read the word, press space bar as quickly as possible. At the end of each sentence, you will decide how acceptable that sentence is to you on a scale from 1 to 4, where 1 is unacceptable, 2 is slightly unacceptable, 3 is slightly acceptable, and 4 is acceptable. An acceptable sentence is: 1) Any way that you normally speak, 2) Something another native speaker of American English could say, 3) Something you or another native use of American English could text or message someone in everyday life".

The purpose of the practice block was to familiarize participants with the self-paced reading procedure (Figure 4-1) and troubleshoot any technical issues that result from an online study. All participants read and rated the same seven sentences drawn from two constructions: Dialectal Adverb (e.g., preverbal *quick* construction from section 1.2.2; Congruous: We could *quickly* cook some dinner.; Dialectal: We could *quick* cook some dinner. ; Incongruous: We could *quicker* cook some dinner.). and Lexical Semantic (see Appendix D for the practice block stimuli). Participants were not provided any feedback on their performance throughout the practice block or main experiment.

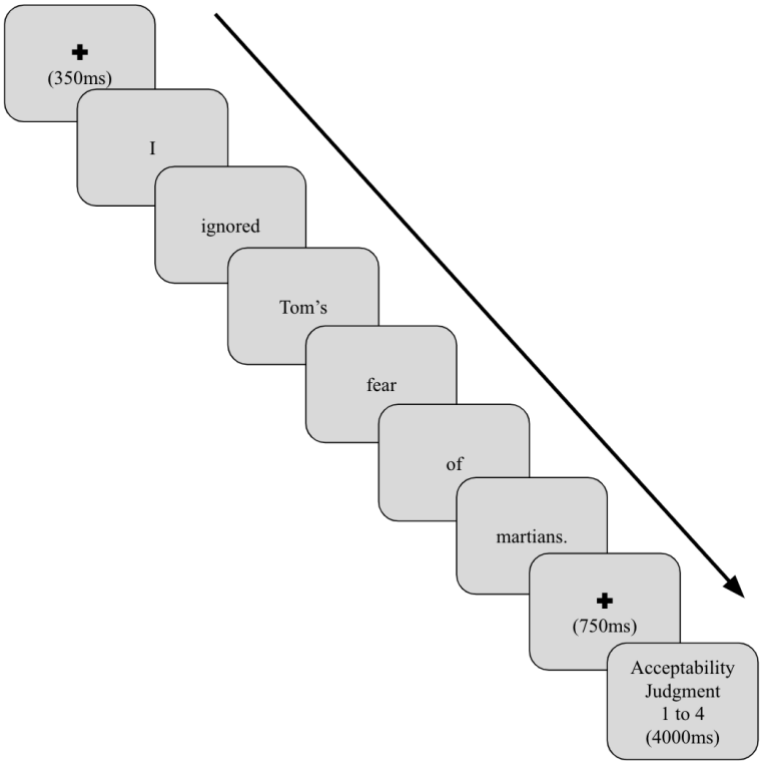


*Figure 4-1: Experimental Procedure using Self-Paced Reading Paradigm.* Each word of the sentence is presented in isolation in the middle of the screen. Participants pressed the space bar to immediately advance to the next word of the sentence. A 4-point Likert-scale (1 unacceptable to 4 acceptable) was used to rate the acceptability of the previous sentence.

As shown in Figure 4-1, each trial began with a fixation cross. Participants read sentences word-by-word—advancing immediately to the next word using the space bar (Just et al., 1982)—in a self-paced reading protocol administered via Gorilla.sc. The end of the sentence was

indicated with a period on the last word. At the end of each trial, participants rated sentence acceptability on a four-point Likert-scale (Dröge et al., 2016). The acceptability judgment prompt remained on the screen for maximally 4 seconds or until the participant entered their response. Breaks followed each block of the experiment allowing participants to rest their eyes, drink some water, and/or mentally reset before proceeding to the next block.

At the end of the experiment, participants completed an open response sociolinguistic survey. Participants answered questions about their demographic information, language background, and locations they have lived. See Appendix E for the sociolinguistic survey.

### *4.2.3 Stimuli*

These conditions and experimental blocking are identical to the EEG study in Chapter 2. 520 items were drawn from four conditions (Table 4-1): Gender Reflexive (60), Complementizer (80), Subject-Verb Agreement (60), and Lexical Semantic Sentences (60).

*Table 4-1: Stimuli Design and Acceptability Predictions.* The target word is underlined and the target word +1 is italicized for the spillover region. Acceptability predictions: ✓ = acceptable, * = unacceptable, ✓/* = may be acceptable or unacceptable.

| Condition | Preference | Total # of Stimuli | Example Stimuli |
|---|---|---|---|
| Target: Complementizer (CP) | Congruous | 40 | ✓ The observation <u>that</u> *six* monkeys are swinging delighted the child. |
| | Incongruous | 40 | ✓/* The observation <u>these</u> *six* monkeys are swinging delighted the child. |
| Control: Subject-Verb Agreement (SVA) | Congruous | 30 | ✓ The cats <u>meow</u> *by* the window watching the birds. |
| | Incongruous | 30 | * The cats <u>meows</u> *by* the window watching the birds. |
| Filler: Gender Reflexive (GR) | Congruous | 30 | ✓ The calm bride prepared <u>herself</u> *for* the wedding. |
| | Incongruous | 30 | * The calm bride prepared <u>himself</u> *for* the wedding. |
| Filler: Lexical Semantic (SS) | Congruous | 30 | ✓ The gardeners trimmed the <u>shrubs</u> *last* Monday. |
| | Incongruous | 30 | * The gardeners trimmed the <u>purse</u> *last* Monday. |

The sentences were separated into two lists and randomized across participants such that each participant read and rated the acceptability of 260 American English sentences to avoid repetition effects. The experiment was split into 10 randomized blocks each containing 26 sentences and untimed breaks followed each block. As shown in Table 4-1 each block consisted of a fixed number of sentences: 8 from the Target and 6 sentences each from the Subject-Verb Agreement Controls, Lexical Semantic Fillers, and Gender Reflexive Fillers.

### *4.2.4 Predictions*

Table 4-2 indicates predictions for both RTs at Target Word +1 and acceptability judgments for each condition. In general, incongruous preferences are predicted to be read slower than the congruous preference at the target word +1, where SPR reading differences are most likely.

*Table 4-2: Online and Offline Predictions in all Conditions.* The example stimuli have the target word underlined and the target word +1 is italicized. Acceptability judgments are on a 4-point Likert-Scale (1 *unacceptable* to 4 *acceptable*). The RT predictions are informed by the self-paced reading literature and EEG work by Tanner (2019), while the acceptability judgment predictions are informed by the acceptability judgment results in Chapter 2.

| **Conditions** | **Preference** | **Example Stimuli** | **Predictions** | |
|---|---|---|---|---|
| | | | **RT (Target Word +1)** | **Acceptability Judgment (1-4)** |
| Complementizer (CP) | Congruous | The observation <u>that</u> *six* monkeys are swinging delighted the child. | Baseline RT (faster than Incongruous) | 3-4 |
| | Incongruous | The observation <u>these</u> *six* monkeys are swinging delighted the child. | Increase in RT compared to Congruous | 1-4 |
| Gender Reflexive (GR) | Congruous | The calm bride prepared <u>herself</u> *for* the wedding. | Baseline RT (faster than Incongruous) | 3-4 |
| | Incongruous | The calm bride prepared <u>himself</u> *for* the wedding. | Increase in RT compared to Congruous | 1-2 |
| Lexical Semantic (SS) | Congruous | The gardeners trimmed the <u>shrubs</u> *last* Monday. | Baseline RT (faster than Incongruous) | 3-4 |
| | Incongruous | The gardeners trimmed the <u>purse</u> *last* Monday. | Increase in RT compared to Congruous | 1-2 |
| Subject-Verb Agreement (SVA) | Congruous | The cats <u>meow</u> *by* the window watching the birds. | Baseline RT (faster than Incongruous) | 3-4 |
| | Incongruous | The cats <u>meows</u> *by* the window watching the birds. | Increase in RT compared to Congruous | 1-4 |

The acceptability judgment predictions in Table 4-2 are informed by the acceptability judgment results from Chapter 2. The congruous preferences in all conditions are expected to be rated as acceptable—either 3 or 4, while the incongruous preferences may be rated more categorical or gradient. The gender reflexive and lexical semantic incongruous preferences are predicted to be more categorical in the acceptability judgments. These stimuli are predicted to be unacceptable rated at either 1 or 2. The complementizer and subject-verb agreement incongruous

preferences are predicted to display gradient acceptability judgments using the entire 4-point Likert-Scale.

To understand the relationship between acceptability judgments and RTs, imagine a graph with acceptability across the x-axis and RT on the y-axis. If the relationship between acceptability and RT is proportional in conditions demonstrating gradient acceptability, then I would predict a negative relationship. The lower acceptability ratings are expected to occur with the slowest RTs (e.g., lower acceptability indicates slower RTs). As the acceptability ratings increase, I expect the RTs to decrease (e.g., higher acceptability indicates faster RTs). This would lead to a negative proportional relationship between the online and offline measures.

### *4.2.5 Data Exclusion Criteria and Data Analysis*

23 participants' datasets were excluded by the following criteria for having average RTs under 300 ms and greater than 1500 ms or for participants missing more than ten percent of their acceptability judgment responses. The remaining 57 participants included in the analysis ranged in age from 18-67 years old (mean age= 32.1). Participants identified their genders as 2 nonbinary, 25 female, 29 male, and 1 did not wish to disclose.

Individual RTs less than 200 ms or greater than 2000 ms were excluded from the remaining participants, as were timed-out acceptability judgments. RT data were adjusted to account for varying position of the target word. This meant that the Relative Word (RelWord) data had the target word equal to 0 and the other words were numbered relative to their distance from the target. RTs and acceptability judgments were then averaged for each participant and condition.

To assess the relationship between acceptability judgments and RTs, I fit a linear mixed effects model with RT as a dependent measure and acceptability, condition, and their interaction

as a fixed effect. I included subjects as a random intercept and random slopes for their acceptability, condition, and their interaction. Priors were set using the defaults of the brms package (version 2.17.0) in R.

## 4.3 Results

The results section will be structured by first examining these measures before the analysis of both measures simultaneously. First, I will begin with a discussion of the offline acceptability judgment responses. Next, the online RTs will be examined. Then, the separate RT and acceptability judgment graphs will be combined to visualize the relationship between these two measures. The relationship will be further examined through a linear mixed effects model.

### *4.3.1 Offline Behavioral Acceptability Judgment Responses*

Figure 4-2 shows the average and by-subject acceptability judgments for each preference in each condition. The average acceptability judgments for the congruous preferences are rated between 3 and 4 in all four conditions, while the average acceptability judgments for the incongruous preferences vary across the four conditions.

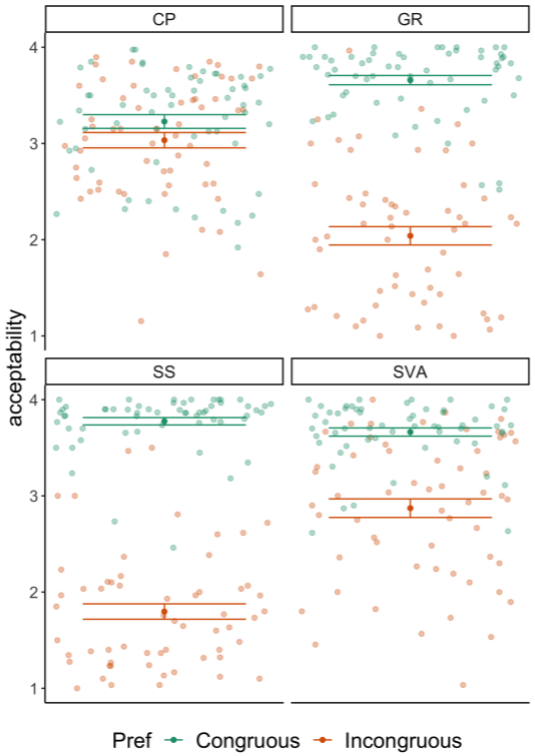


*Figure 4-2: Average Acceptability Judgment Results By-subject.* The x-axis is a placeholder to show both preferences (green= congruous, orange=incongruous) in each condition: CP, GR, SS, and SVA. Acceptability is listed across the y-axis, where 1 is *unacceptable* and 4 is *acceptable.* The bolded dots represent the average acceptability of that preference in one condition for all participants, and the standard error is shown by the error bars for each of the averages. The transparent dots represent the by-subject average for both preferences in each condition. [n=57].

The lexical semantic and gender reflexive conditions appear to be categorical, while the complementizer and subject-verb agreement conditions appear to be gradient. The incongruous Gender Reflexive and Lexical Semantic average acceptability judgments are rated just below 2. The participants rate the Lexical Semantic congruous preference between 3-4, while the incongruous preference is rated between 1-2. The congruous Gender Reflexive preference is rated similarly to the congruous Lexical Semantic condition at 3-4. The incongruous Gender Reflexive preference shows more gradience in the incongruous preference[27] with the by-subject averages spanning 1-3.

[27] This could be due to a larger discussion around gender since 2019 when the EEG study was completed.

The incongruous complementizer average acceptability judgment is lower than the average congruous acceptability judgment at about 3. The by-subject acceptability averages in the complementizer condition show that both the incongruous and congruous preferences are rated between 2-4, which replicates the average acceptability results in Chapter 2. The average acceptability ratings from subject-verb agreement condition behaved as expected. The congruous preference is rated between 3-4, while the incongruous preference has by-subject averages at all four Likert-Scale points.

### *4.3.2 Online Measure: RTs*

Figure 4-3 shows the averaged RTs in milliseconds from two words before and two words after the target word in all four conditions.

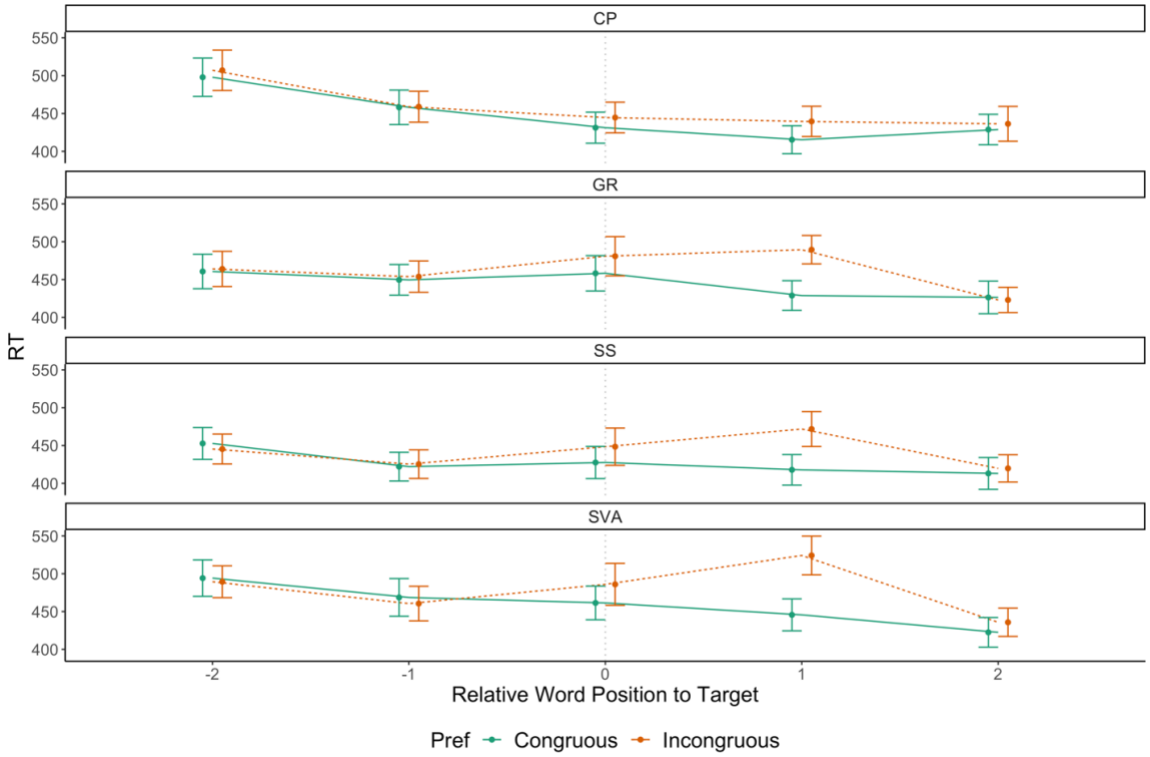


*Figure 4-3: Averaged Self-Paced Reading Time Results.* The averaged RTs in milliseconds (y-axis) are shown by their position relative to the target word (x-axis), which is located at 0 and highlighted by the vertical grey dotted lines. Each horizontal row represents one of the four conditions in the experiment: Complementizer (CP), Gender Reflexive (GR), Lexical Semantic Sentences (SS), and Subject-Verb Agreement (SVA). The lines are colored green and orange to represent the preference of each condition: congruous and incongruous. [n=57].

Overall, the incongruous preferences are read slower in the word immediately after the target word (Target+1) than the baseline congruous preferences across all conditions. For the complementizer condition, this difference is small with significant overlap in the standard errors of the mean RT.

### *4.3.3 Relationship of the Offline and Online Measures*

#### 4.3.3.1 Visual Relationship between Reading Time and Acceptability

Figure 4-4 combines the averaged self-paced RT and average acceptability judgment results across the four conditions. The y-axis on the average acceptability judgment graph has been transformed by -1 to better visualize the relationship between the average RT and average acceptability judgments.

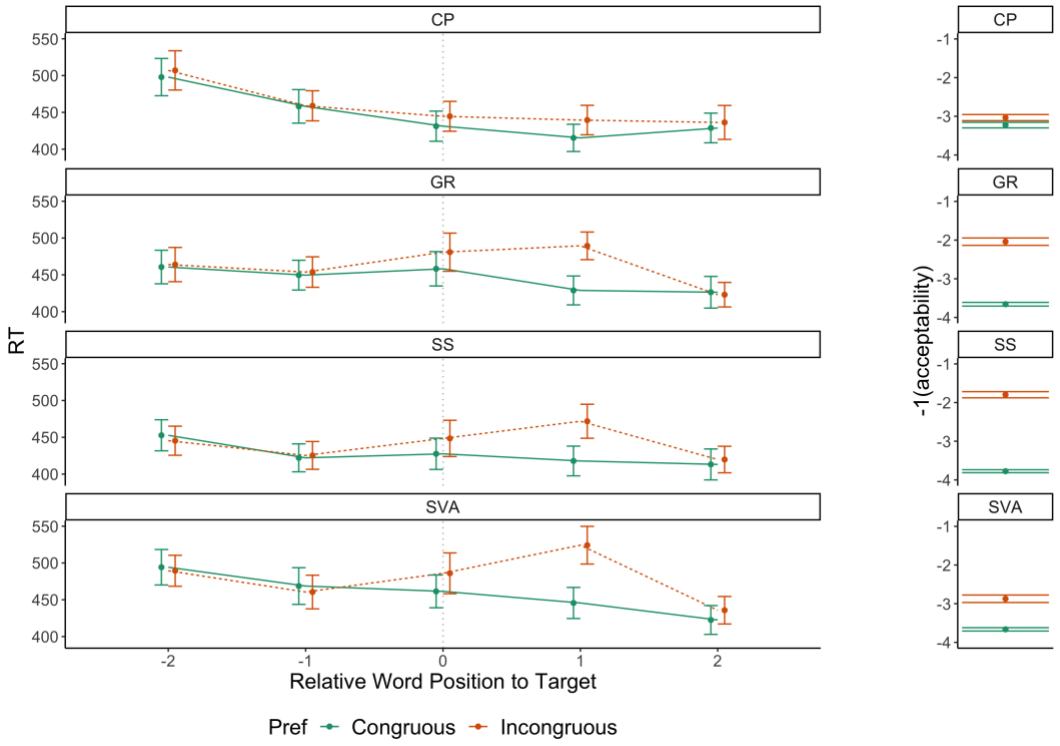


*Figure 4-4: Comparison of the Averaged Self-Paced Reading Time (left) and Averaged Acceptability Judgment (right) Results.* Left: reprint of Figure *4-3*. Right: The averaged acceptability judgment responses (Figure *4-2*) have been multiplied by -1 to better visualize the relationship between RTs and acceptability. -1 is unacceptable and -4 is acceptable. The dots represent the average acceptability of that preference in one condition for all participants, and the standard error is shown by the error bars for each of the averages. [n=57; Preferences: green= congruous, orange= incongruous; Conditions CP, GR, SS, and SVA].

Slower RTs at Target+1 are reflected alongside lower acceptability ratings in the incongruous preferences across all four conditions. The baseline congruous preferences contain the faster RTs at Target+1, as well as higher acceptability judgment ratings. The preferences within the Complementizer condition show similar averages in both RTs and acceptability judgments, in that the standard errors of the mean RT have significant overlap and the standard errors of the mean acceptability are close to each other. This could indicate that both preferences are equally acceptable leading to the processing of both preferences being equally easy resulting in similar RTs.

#### 4.3.3.2 Results for the Linear Mixed Model

Figure 4-5 demonstrates the relationship between acceptability and log-adjusted RTs for every trial for each participant in all four conditions.

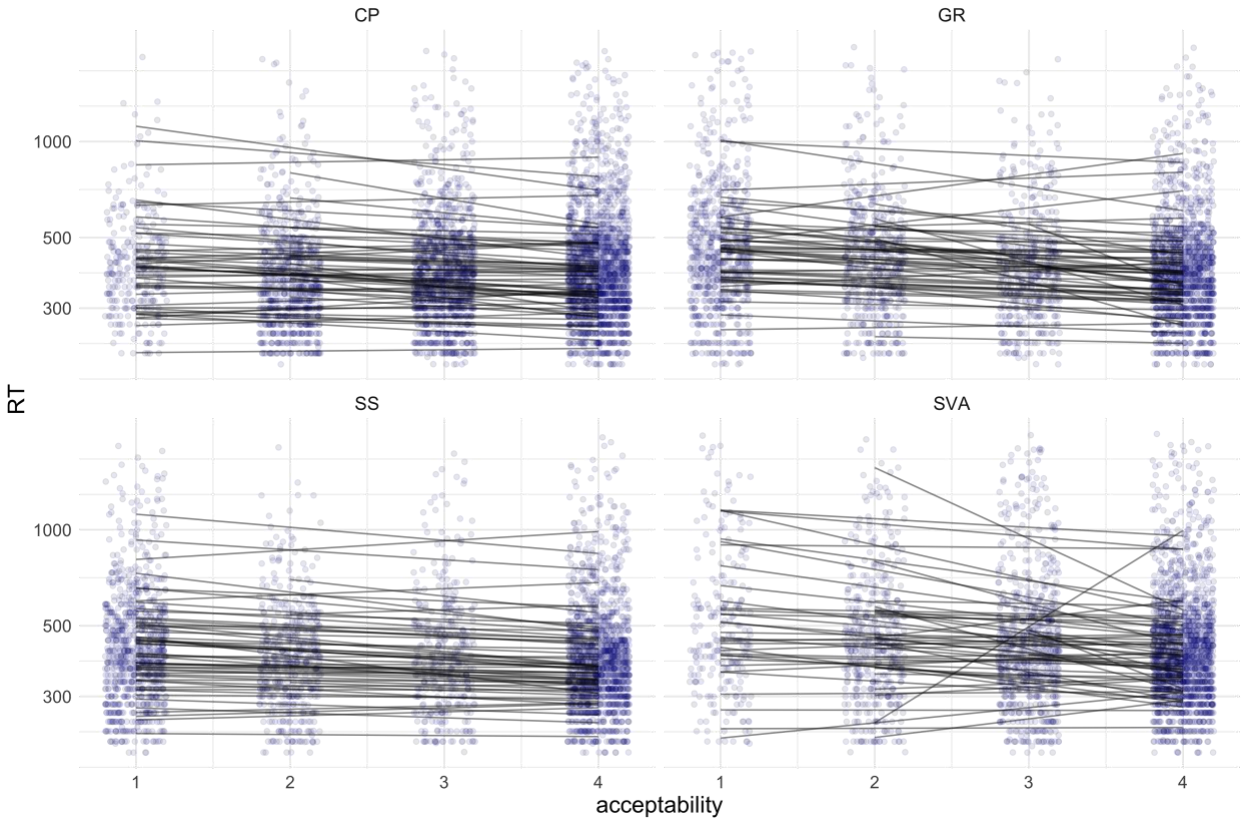


*Figure 4-5: Relationship between Acceptability and Reading Time*. Each panel represents one of the four conditions in the experiment: Complementizer (CP), Gender Reflexive (GR), Lexical Semantic Sentences (SS), and Subject-Verb Agreement (SVA). Each panel shows acceptability judgments across the x-axis and the log-adjusted RTs across the y-axis. The blue dots represent the acceptability judgments & RTs for every trial for each participant. The black lines represent the linear trend for each participant. [n=58].

Across the four panels in Figure 4-5, participants generally show a small negative relationship in that the RTs for unacceptable ratings tend to be higher than the RTs for acceptable ratings in all four conditions. However, this visual relationship is not uniform for every participant, as some participants trends tend to be flat or opposite of the expected relationship (e.g., in the subject-verb agreement panel, the acceptability judgment 2 is read faster than acceptability judgment 4 for one participant).

A linear mixed effects model was performed and the statistical summary of the linear trend in each condition is shown in Table 4-3.

*Table 4-3: Statistical Summary of the Linear Trend per Condition*. The acceptability trend, lower HPD, and upper HPD are slopes across the four conditions in the experiment: Complementizer (CP), Gender Reflexive (GR), Lexical Semantic Sentences (SS), and Subject-Verb Agreement (SVA).

| Condition | Acceptability Trend | Lower HPD | Upper HPD |
|---|---|---|---|
| Complementizer | -0.0312 | -0.0439 | -0.0195 |
| Subject-Verb Agreement | -0.0496 | -0.0665 | -0.0306 |
| Lexical Semantic | -0.0356 | -0.0468 | -0.0249 |
| Gender Reflexive | -0.0704 | -0.0980 | -0.0411 |

All four conditions show a negative acceptability judgment slope, as well as negative slopes for both the lower and upper limits of the credible intervals. This suggests that the relationship between acceptability judgments and RTs is negative (e.g., RTs decrease as acceptability judgments increase). The acceptability judgment trend across all four conditions is statistically reliable, but it is not large. This means that the effect is highly variable across the participants (as shown in Figure 4-5), but on average, if I were to collect data from a new set of participants, the statistics suggest that the average would result in a negative slope for the new set of participants.

## 4.4 Discussion

The experimental methodology from Chapter 2 was altered to include a self-paced reading protocol to collect online RT measures and offline acceptability judgments. In using identical stimuli as the EEG experiment, the acceptability judgment results from the EEG experiment and this experiment were similar. However, the incongruous preferences exhibited differences in the online measures between neural responses and RTs.

The complementizer and the subject-verb agreement conditions were initially selected because the complementizer condition is argued to be gradient from the literature, whereas the literature predicts that the subject-verb agreement condition should be categorical. The results of

the experiment are not in line with the predictions from the literature, as the incongruous subject-verb agreement conditions shows gradience in acceptability.

The online RT measures in the complementizer condition show similar average RTs for both preferences with overlapping error bars at Target+1. The by-subject averages in the complementizer condition are clustered between 2-4, with most of the averages clustering between 3-4 without overlapping error bars. The close averages for both preferences in the online and offline measures could indicate that both are equally acceptable, which leads to the processing of the CP condition being equally as easy, thereby resulting in similar RTs.

The online RT measures in the subject-verb agreement condition show distinct average RTs for both preferences at Target+1. The higher than expected acceptability judgments for the incongruous preference in the subject-verb agreement condition matches previous data from Tanner (2019) and the results from Chapter 2. Due to the experimental procedure in Figure 4-1, the violation in the lexical subject-verb agreement seems likely to be repaired, forgotten, or missed in these sentences leading them to be judged more acceptable in the offline measure.

In comparing the relationship between RTs and acceptability among these two conditions, subjects showed a negative relationship in that RTs for unacceptable ratings tend to be slower than the RTs for acceptable ratings. The model suggests that the Subject-Verb Agreement condition has a slightly stronger acceptability trend than the Complementizer condition. While not uniform among participants, the relationship between online and offline measures in these two conditions is proportional albeit not very large.

As mentioned in section 4.1, Tanner (2019) had found reliable online ERP effect in using the RSVP and SPR protocols. The results from Chapter 2 and this study find some key differences in the correlation between the two online measures (e.g., ERP amplitudes and RTs)

and acceptability judgments. Both analyses observe negative trends when correlating these measures; however, the reliability of the trends differs. The analysis in Chapter 2 observes a negative trend for Subject-Verb Agreement, but the other conditions did not produce a reliable trend. In the SVA condition, more acceptable sentences yield less positive ERP amplitudes. In contrast, all four conditions in this current study observe a reliable negative trend, which indicates that more acceptable sentences yield faster RTs. With respect to online measures, RTs produce a stronger correlation with acceptability judgments than ERP amplitudes for these conditions. This suggests that the proportional relationship between online and offline measures is more reliably represented for RTs as the online measure than ERP amplitudes.

### 4.5 Summary of the Study

The current study tests whether gradient acceptability judgments to a target condition show proportional gradience in online RTs using an online self-paced reading task. Participants read and rated the acceptability of each sentence across four conditions: complementizer, subject-verb agreement, gender reflexive, and lexical semantic. A linear mixed effects model was fit with RT as a dependent measure and acceptability, condition, and their interaction as a fixed effect. The model included subjects as a random intercept and random slopes for their acceptability, condition, and their interaction. Given the high variability in the relationship between RTs and acceptability judgments across the participants, the results show a small, but statistically reliable negative relationship across all four conditions. The overall results suggest a proportional relationship between offline and online measures across all conditions, in that as acceptability judgments increase, RTs decrease.

# Chapter 5 Exploring the Effect of a Dialectal Condition in Online Reading Time Measures and Offline Acceptability Measures Using Self-Paced Reading (SPR)

This experiment builds on the results from Chapter 2 and Chapter 4 by examining how language experience and familiarity affects the link between offline acceptability judgments and online RTs using a dialectal construction as the target. The dialectal construction under investigation is the needs+past participle construction (e.g., The bathroom floor *needs mopped* on Thursday), which is used in Pennsylvania, Ohio, Indiana, Iowa, Kansas, Nebraska, and South Dakota (Kaschak & Glenberg, 2004; Murray & Simon, 2006). Previous studies that pair multiple measures with dialect-specific constructions have analyzed these measures separately (García, 2017; Zaharchuk et al., 2021). These studies have not shown a consistent or fixed relationship between offline and online measures. The current study probes if gradient acceptability judgments to a target dialectal condition show proportional gradience in online reading when modulated by language exposure. This modulation is needed to assess whether the negative relationship between RT and acceptability judgments remains or changes within the dialectal condition.

## 5.1 Literature Review

First, section 5.1.1 will review the impact of speaker identity and dialectal familiarity on the online measures. Next, section 5.1.2 examines the impact of awareness and dialectal familiarity on both the online and offline measures. This section builds on the discussion of García in chapter 2 about the relationship between online and offline measures and discusses

how awareness may be at play. Section 5.1.3 then examines the literature behind the needs+past participle construction. This leads to the current research question (section 5.1.4) and the conditions used to test this research question (section 5.1.5).

### *5.1.1 The Impact of Speaker Identity and Familiarity on Online Measures*

Several sociolinguistic and psycholinguistic studies have shown that speaker identity and dialectal experience modulate online processing (Table 5-1).

*Table 5-1: Sociolinguistic Evidence of Online Measure Modulation.*

| Authors | Modulates Online Measure | Sociolinguistic Features within the Experiments | Evidence for Modulation of Online Measure |
|---|---|---|---|
| Weissler & Brennan (2020) and Weissler (2021) | Speaker identity and language variety | Auxiliary *be* usage in African American English (AAE) and Standardized American English (SdAE) | Speaker identity and language variety modulated the P600 response (e.g., no neural differences in grammatical and ungrammatical forms for minoritized variety) |
| Hanulíková et al. (2012) | Speaker Identity | Gender Agreement constructions with Dutch and Turkish-accented Dutch speakers | P600 is modulated by speaker identity (e.g., L2 accented speaker is expected to make mistakes therefore no P600 response) |
| Van Berkum et al. (2008) | Speaker Identity | Stimuli based on stereotypes about the speaker based on gender, age, socioeconomic class | N400 is modulated when speaker identity conflicts with sentential context (e.g., "I like wine before bed" spoken by a child) |
| Squires (2014a) | Dialectal experience/ familiarity | White and Black Readers with non-standard *don't* and the uncommon *doesn't* constructions | Knowledge of agreement variability modulated the RTs of the Black readers (e.g., nonstandard variety read the slowest as opposed to the uncommon construction). |
| García (2017) | Dialectal experience/ familiarity | Omitted -S in Mono-dialectal speakers of MUSE and Bi-dialectal speakers of MUSE and AAE | P600 was modulated based on dialectal experience (e.g., Bi-dialectal speakers showed no P600 response to the omitted -S condition) |
| Zaharchuk et al. (2021) | Dialectal experience/ familiarity | Double modal in familiar (Southern) and unfamiliar (Nonsouthern) speakers | Familiarity did not modulate the ERP responses as both groups showed ELAN and P600 responses |

In her dissertation, Weissler (2021) conducted two auditory EEG experiments to test whether listeners are taking the speaker's identity and language variety into account (Following

research from Hanulíková et al., 2012 and Van Berkum et al., 2008). Two male speakers from the Midwest were recorded speaking three different constructions: Auxiliary present, Auxiliary absent, and ungrammatical variant. One speaker was a bidialectal Black speaker of AAE and SdAE, in which both auxiliary constructions are grammatical in his dialects. The other speaker was a mono-dialectal white speaker of SdAE and only the auxiliary present condition is grammatical in his dialect. Results indicated that AAE and SdAE are processed differently when produced by a Black bidialectal speaker and a white speaker. There were no P600 responses across all conditions for the bidialectal speaker, while the auxiliary absent and ungrammatical variant showed P600 responses. These results suggest that listeners take the speaker's identity and language variety into account during online processing.

Squires (2014a) provides additional evidence of dialectal experience affecting the online processing based on reader identity. Four constructions were created for a self-paced reading study that varied as a function of subject agreement with the negated *do* verb (e.g., standard PL+*don't/* SG+*doesn't*, nonstandard SG+*don't*, and uncommon PL+*doesn't* constructions). Squires demonstrates that dialectal experience impacts the processing of the nonstandard *don't* variant. White readers showed significant differences in RTs across all constructions with RTs ranging from fastest in the standard to slowest in the uncommon constructions. Black readers also had the slowest RTs in the uncommon construction, but RTs for the standard constructions and the nonstandard construction were a similar speed for the Black readers. Squires attributes this processing difference is likely attributed to the Black readers being exposed to the nonstandard variant in their own dialects, while the white readers lack this exposure.

### *5.1.2 The Impact of Awareness and Familiarity on Online and Offline Measures*

Building on her previous work, Squires (2016) tested whether participants perceived and noticed grammatical differences in the processing of the same grammatically variant constructions from above and grouped them by awareness based on their answer to a post-experiment question. Squires defines awareness as the emergence "from aggregated experiences of in-the-moment noticing of linguistic differences – and coming to understand them as linguistically and socially meaningful" (2016, p. 82). Grouping by awareness modulated the RTs, such that the aware group read the nonstandard and uncommon constructions significantly slower than the standard construction. The unaware group only read the uncommon construction significantly slower than the standard construction. These results suggest that level of awareness modulates online processing. Crucially, this study provides evidence that grammatical variants that are subject to conscious awareness are processed differently, that is the nonstandard variants (e.g., nonstandard and uncommon) are process slower than the same variants that are not subject to conscious awareness.

In an effort to combine ERP and sociolinguistic measures, Zaharchuk et al. (2021) conducted an auditory ERP study and a post-experiment acceptability judgment task using the double model syntactic construction (e.g., might could). The stimuli consisted of single modal and attested double modal constructions, and the acceptability judgment task also included unattested double modal constructions. The familiar (southern) and unfamiliar (unmarked) groups were created based on each participants' double modal exposure, double modal familiarity, and geographic backgrounds. The online ERP results for both the familiar and unfamiliar groups showed an early anterior negativity and P600, which suggests that dialect experience does not affect online processing. In contrast, the offline acceptability judgments

were modulated by dialect familiarity. Southern participants were more sensitive to all three modal constructions than the unfamiliar participants. The mismatch between these online and offline measures suggests that dialect familiarity only modulates offline measures.

### *5.1.3 Dialectal Needs Construction in the Literature*

Another construction similar to the double modal construction is the needs+past participle construction from the United States midland dialect. This construction has reported use in Pennsylvania, Ohio, Indiana, Iowa, Kansas, Nebraska, and South Dakota (Kaschak & Glenberg, 2004; Murray & Simon, 2006). Murray and Simon explain that the dialectal needs construction is stigmatized, as it is "currently rejected as ungrammatical by several times more Midland residents than accept or use them (although generally not by users themselves)" (2006, p. 29).Although a large portion of Midland residents find the construction unacceptable, it is unclear if this language attitude will affect online measures (e.g., RTs for the needs+past participle construction as slow as an ungrammatical construction) or usage in informal speech, as language users may not realize they use the construction that they report is unacceptable (see Chapter 1.2.2 for previously discussed examples).

The construction is typically spoken, but the construction is now being written more often (but still infrequently) on the social media platform Twitter (Strelluf, 2020). Strelluf conducted a corpus study on Twitter analyzing the written use of needs+past participle and found that most of the tweets containing the construction were about sports. The tweets using the needs+past participle constructions were primarily from users from Ohio (except for the Cleveland area) and mostly western Pennsylvania (the Pittsburgh area). Written usage of this dialectal construction provides evidence of this grammatical variant, which is especially important as unconscious usage and reported acceptability often vary (as discussed in section 1.2.2). Specifically, Labov et

al. (2005) compare the needs+past participle construction to positive *anymore*: "needs +PPtc are subject to the same uncertainties as positive *anymore*, in that conscious recognition falls short of spontaneous speech" (p. 296).

Previous research on the needs+past participle construction has shown that participants without prior linguistic experience adapted to the unfamiliar construction over the course of the self-paced reading experiment (Fraundorf & Jaeger, 2016; Kaschak, 2006; Kaschak & Glenberg, 2004). The repeated exposure of the unfamiliar construction resulted in faster RTs, such that participants without prior linguistic experience read more like experienced participants with the needs+past participle construction. Additionally, Fraundorf and Jaeger (2016) found that perceiving the needs+past participle construction as unusual contributed to slower RTs than those who did not perceive the construction. This finding suggests that linguistic experience and familiarity with the needs+past participle construction could affect the online RTs, which in turn could lead to gradience in an online measure.

### *5.1.4 Research Question*

The mismatch in online and offline measures observed by García (2017) and another mismatch observed by Zaharchuk et al. (2021) both lead to an important question about what the relationship between online and offline measures are when both measures may be sensitive to language experience. The results from Zaharchuk et al. (2021) suggest that the online and offline measures may be differently affected when modulated by language experience and familiarity. One reason for this difference could that the "online data may be more sensitive to the immediate linguistic environment than offline data" (Zaharchuk et al., 2021, p. 11). As both studies collected the offline acceptability judgments after the EEG portion of the experiment, this leaves open the possibility that the collection of the online measures and acceptability judgments within

the same trial may lead to new insights on how language experience modulates either online or offline measure.

As discussed at the end of Chapter 1.2.2, studies from Labov (1996) Hildebrandt (2017) suggested that language users vary on their awareness regarding their own dialects and then report that they find their own dialectal construction unacceptable causing gradience in acceptability for dialectal constructions. This experiment aims to test how language experience[28] modulates the relationship between the online RTs and offline acceptability judgments when using the dialectal construction needs+past participle. Specifically: do gradient acceptability judgments to a target dialectal condition show proportional gradience in online RTs when modulated by language exposure[29]? The trend in Chapter 4 pointed to a negative correlation between acceptability judgments and RTs, such that lower acceptability judgments are correlated with slower RTs.

### *5.1.5 Conditions*

#### 5.1.5.1 Needs + Past Participle Construction

The needs+past participle construction serves as the dialectal variant that may differ in the online and offline measures based on the language experience of the participant. Following previous research, this condition was constructed of three minimally different variants: congruous, incongruous, and dialectal (Squires, 2016; Weissler, 2021; Weissler & Brennan, 2020; Zaharchuk et al., 2021). The congruous variant (e.g., The bathroom floor needs mopping on Thursday) is acceptable in all American English dialects and the incongruous variant (e.g.,

[28] I would argue that language attitudes are included in the geographical-based language experience.
[29] Dialectal constructions may also be stigmatized such that negative language attitudes could affect the online and offline measures. This chapter focuses on language exposure, while chapter 6.2.2 explores the effect of awareness and indexing on this set of data.

The bathroom floor needs mopper on Thursday) is unacceptable in all American English dialects. The dialectal variant (e.g., The bathroom floor needs mopped on Thursday) is acceptable to some speakers of the midland dialect variety but not to speakers of other American English dialects (Murray & Simon, 2006; Squires, 2016). The dialectal variant may be perceived as a syntactic violation for those who do not speak the midland dialect or do not have enough location-based exposure to the construction.

The geographical divide in who uses this construction serves as a test of if a dialectal construction alters the relationship between online and offline measures. Previous studies have shown a mismatch in these measures, such that linguistic experience and familiarity affect the online and offline measures differently (García, 2017; Zaharchuk et al., 2021). The offline acceptability judgments in García mismatched the lack of the P600 response in the bi-dialectal group. Whereas the online ERP effects in Zaharchuk et al. indicated that participant processed the double modal as a violation, but the acceptability judgments were rated acceptable.

#### 5.1.5.2 Phrase Structure Construction

Previous self-paced reading and EEG studies established reliable online (ERP and RTs) effects in sentences containing a phrase structure violation (Neville et al., 1991; Hagoort et al., 1993; Osterhout & Holcomb, 1992; Friederici & Meyer, 2004). The phrase structure violations focused on switching the preposition and noun phrase in the object position (e.g., The children enjoyed Ed's about stories the farm). The phrase structure construction was also selected because changes in word order are included in everyone’s grammar. These syntactic violations are likely very noticeable to participants and so phrase structure violations are less likely to be repaired, forgotten, or missed in sentences (unlike the subject-verb agreement violations), which in turn, will result in more reliability in the offline acceptability judgments.

#### 5.1.5.3 Gender Reflexive & Lexical Semantic Constructions

These conditions were selected as previous self-paced reading and EEG studies established reliable online (ERP and RTs) effects in sentences containing a violation (Carreiras et al., 1996; Vincenzi et al., 2003; Tanner, 2019; Kutas & Hillyard, 1980). The gender reflexive violations consisted of a gender disagreement between the reflexive and the subject (e.g., The hungry waitress ordered himself a burger). The lexical semantic violation consisted of an object noun conflicting with the sentential context (e.g., Mike delivered the beard to the wrong apartment). Additionally, these conditions were used in both Chapter 2 and Chapter 4, and they will be used as a sanity check to ensure this experiment is reliable.

## 5.2 Methods

### *5.2.1 Participants*

170 participants were recruited on Prolific across two groups (n=85) based on where the participants were born and their current location. Participants who were born in and/or currently located in Ohio and Pennsylvania were grouped into the +OHPA group[30] and those born in and/or currently located in any other state were grouped into the -OHPA group. These groups were selected to correspond to where the *needs* condition is spoken. The +OHPA group are likely to be familiar with the *needs* condition, while those in the -OHPA group are less likely to be familiar with the *needs* condition. This creates the possibility that one's own grammar or the surrounding linguistic input may affect the online and offline measures of the experiment. Familiarity with the dialectal construction was also assessed using a post-experiment

[30] Four participants were not born in either Ohio or Pennsylvania, but they had all spent extensive periods of time in these states at some point in their lives. All participants had lived in Ohio or Pennsylvania at some point in their lives for long periods of time.

questionnaire (see section 5.2.2 and Appendix G). The participants were compensated at the rate of $9.50 per hour for their participation.

### *5.2.2 Self-Paced Reading and Acceptability Judgment Procedure*

The experimental procedure was nearly identical to Section 4.2.2 apart from the total time of the study, stimuli design in the practice block, the addition of timed breaks, and the sociolinguistic demographic survey. These differences will be highlighted within this section.

The experiment was designed to be completed in 35 minutes, but participants were allowed 100 minutes to complete all three tasks. Participants took an average of 23 minutes to complete the experiment. Timed one-minute breaks were added to this experiment to encourage online participants to finish the experiment in one sitting.

Within the practice block, participants read and rated nine sentences drawn from two constructions: Phrase Structure stimuli that were not used in the main experiment and Dialectal Adverb stimuli (as defined in Section 4.2.2; see Appendix F for experiment specific stimuli). Participants were not provided any feedback on their performance throughout the practice block or main experiment.

At the end of the experiment, participants completed an open response sociolinguistic survey (Appendix G), which included questions about their demographic information, language background, and locations they have lived. Additionally, participants answered questions about their awareness and indexation of the dialectal *needs* condition.

### *5.2.3 Stimuli*

240 items were drawn from four conditions (Table 5-2): Gender Reflexive (60), Needs (60), Phrase Structure (60), and Lexical Semantic Sentences (60). Items for each condition included congruous and incongruous variants.

*Table 5-2: Stimuli Design and General Acceptability Predictions.* The target word is underlined and the target word +1 is italicized for the spillover region. Acceptability predictions: ✓ = acceptable, * = unacceptable, ✓/* = +OHPA likely acceptable and -OHPA likely unacceptable.

| Condition | Preference | Total # of Stimuli | Example Stimuli |
|---|---|---|---|
| Gender Reflexive (GR) | Congruous | 15 | ✓ The hungry waitress ordered herself *a* burger. |
| | Incongruous | 15 | * The hungry waitress ordered himself *a* burger. |
| Needs (N) | Congruous | 15 | ✓ The bathroom floor needs mopping *on* Thursday. |
| | Dialectal | 15 | ✓/* The bathroom floor needs mopped *on* Thursday. |
| | Incongruous | 15 | * The bathroom floor needs mopper *on* Thursday. |
| Phrase Structure (PS) | Congruous | 15 | ✓ The children enjoyed Ed's stories *about* the farm. |
| | Incongruous | 15 | * The children enjoyed Ed's about *stories* the farm. |
| Lexical Semantic (SS) | Congruous | 15 | ✓ Mike delivered the pizza *to* the wrong apartment. |
| | Incongruous | 15 | * Mike delivered the beard *to* the wrong apartment. |

The stimuli were presented across four lists using Latin square design such that each participant read and rated the acceptability of 135 sentences. The experiment was split into 5 blocks each containing 27 sentences (three from each Condition Preference).

### *5.2.4 Predictions*

The linguistic experience in both groups is predicted to be the same for the Gender Reflexive, Needs and Phrase Structure constructions across the congruous and incongruous preferences. Linguistic experience is only expected to modulate processing in the dialectal preference in the Needs condition. In this condition, we expect that the -OHPA group likely has less linguistic experience with the dialect-specific variant. We expect that the +OHPA group

likely has linguistic experience with the dialect-specific variant. These differences are represented in Table 5-3.

The Lexical Semantic condition is different as there is not a syntactic violation, but rather a lexical semantic violation. Across the two groups, the semantic knowledge is the same (e.g., reading *beard* when you expect to read *pizza* would violate expectations) and so there are not any predicted group differences in the congruous and incongruous preferences in this condition.

In general, incongruous preferences are predicted to be read slower than the congruous preference. For dialectal preference RT: Participants who were born and live in Ohio and Pennsylvania are predicted to have RTs closer or identical to the congruous preference—perhaps due to their linguistic experience with the dialectal variant. Participants who were born and live in other states are predicted to have RTs closer or identical to the incongruous preference—perhaps lacking linguistic experience with the dialectal variant.

*Table 5-3: Online and Offline Predictions in all Conditions.* The example stimuli have the target word underlined and the target word +1 is italicized as this area is where RT differences are most likely. Group differences are expected for the needs dialectal stimuli. The Needs condition has two predictions based on language experience[31]: a) is expected for the +OHPA group and b) is expected for the -OHPA group.

| Condition | Preference | Example Stimuli | Predictions | |
|---|---|---|---|---|
| | | | RT For Target Word +1 | Acceptability Judgment (1-4) |
| Gender Reflexive | Congruous | The hungry waitress ordered herself *a* burger. | Baseline RT (faster than Incongruous) | 3-4 |
| | Incongruous | The hungry waitress ordered himself *a* burger. | Increase in RT for all groups compared to Congruous | 1-2 |
| Needs | Congruous | The bathroom floor needs mopping *on* Thursday. | Baseline RT (faster than Incongruous) | 3-4 |
| | Dialectal | The bathroom floor needs mopped *on* Thursday. | a) RT closer or identical to Congruous for those with dialectal experience<br><br>b) RT closer or identical to Incongruous for those without dialectal experience | a) 3-4<br><br>b) 1-2 |
| | Incongruous | The bathroom floor needs mopper *on* Thursday. | Increase in RT for all groups compared to Congruous | 1-2 |
| Phrase Structure | Congruous | The children enjoyed Ed's stories *about* the farm. | Baseline RT (faster than Incongruous) | 3-4 |
| | Incongruous | The children enjoyed Ed's about *stories* the farm. | Increase in RT for all groups compared to Congruous | 1-2 |
| Lexical Semantic | Congruous | Mike delivered the pizza *to* the wrong apartment. | Baseline RT (faster than Incongruous) | 3-4 |
| | Incongruous | Mike delivered the beard *to* the wrong apartment. | Increase in RT for all groups compared to Congruous | 1-2 |

The acceptability judgment predictions in Table 5-3 are informed by the acceptability judgment results from Chapter 2, Chapter 4, and Section 5.1.5. The congruous preferences in all

[31] These predictions were made based on language experience, which is grouped based on birth location and current location. However, language attitudes and stigmatization of dialectal needs would not align with these predictions. This is outside the scope of the current chapter, but awareness and indexing will be visually explored in chapter 6.2.2.

conditions are expected to be rated as acceptable as either 3 or 4. The incongruous preferences in all conditions are expected to be rated as unacceptable as either 1 or 2. The dialectal preference predictions are based on the location-based groupings. The +OHPA group is expected to have linguistic experience of this dialectal preference and they will likely rate the condition as acceptable (3 or 4), while the -OHPA group is expected to lack linguistic experience of the dialectal preference. This group is predicted to rate the dialectal preference as unacceptable (1 or 2). However, we could expect gradience in the dialectal preference in the both the +OHPA and -OHPA groups, if the language exposure to the dialectal preference varies within group (e.g., if -OHPA participant has family living in the midland area that uses the dialect, but they themselves were born and live elsewhere in the United States).

To understand the relationship between acceptability judgments and RTs, imagine a graph with acceptability across the x-axis and RT on the y-axis. If the relationship between acceptability and RT is proportional in conditions demonstrating gradient acceptability, then I would predict a negative relationship. The lower acceptability ratings are expected to occur with the slowest RTs (e.g., lower acceptability indicates slower RTs). As the acceptability ratings increase, I expect the RTs to decrease (e.g., higher acceptability indicates faster RTs). I expect that this RT and acceptability judgment relationship would hold regardless of the +/- OHPA grouping because regardless of their grouping, the RT and acceptability judgment predictions align.

### *5.2.5 Data Exclusion Criteria and Data Analysis*

Twenty-three participants' datasets were excluded by the following criteria for having average RTs under 300 ms or greater than 1500 ms or for participants missing more than ten

percent of their acceptability judgment responses. Table 5-4 shows the group demographic information of the participants included in the analysis.

*Table 5-4: Group Demographic Information of the Included Prolific Participants.* +OHPA indicates participants who were born in Ohio or Pennsylvania or who currently live in Ohio or Pennsylvania. -OHPA indicates participants who were born in any other state and who currently live other states.

| Group | Participants | Gender Identity | Age Range | Average Age |
|---|---|---|---|---|
| +OHPA | 73 | 56 Female<br>15 Male<br>2 Additional Gender Descriptors[32] | 18 to 61 | 30.6 |
| -OHPA | 74 | 38 Female<br>32 Male<br>3 Additional Gender Descriptors[32] | 18 to 64 | 32 |

Individual RTs less than 200 ms or greater than 2000 ms were excluded from the remaining participants, as were timed-out acceptability judgments. RT data were adjusted to account for varying position of the target word. RTs and acceptability judgments were then averaged by aligning to the target word in each sentence for each participant and condition.

I fit two linear mixed effects model to first assess the relationship between acceptability judgments and RTs and then to assess how group would impact that relationship. I fit the first model with RT as a dependent measure and acceptability, condition, and their interaction as a fixed effect. I fit the second model with RT as a dependent measure and acceptability, condition, group, and their interaction as a fixed effect. For both models, I included subjects as a random intercept and random slopes for their acceptability, condition, and their interaction. Priors were set using the defaults of the brms package (version 2.17.0) in R.

[32] Participants also described their gender as gender queer, nonbinary, heterosexual, cishet, and genderfluid.

## 5.3 Results

First, I will begin with a discussion of the offline acceptability judgment responses across the +OHPA and -OHPA groups. Next, the online RTs will be examined. Then, the RT and acceptability judgment graphs will be combined to visualize the relationship between these two measures. This relationship will be further examined through two linear mixed effects models to determine the relationship between acceptability judgments and RTs and if group affects this relationship.

Lastly the data will be grouped according to qualitative answers from the sociolinguistic survey regarding awareness of self, awareness of others, and the top three indexing answers to see how awareness and indexing affect the relationship between acceptability judgments and RTs.

### *5.3.1 Offline Behavioral Acceptability Judgment Responses*

Figure 5-1 shows the grand average and by-subject average acceptability judgments results for each preference in each condition across the two groups. The by-subject average acceptability judgments are represented by the transparent dots in the graph. The graph is a useful visualization to see the categorical versus gradient nature of the preferences in each condition.

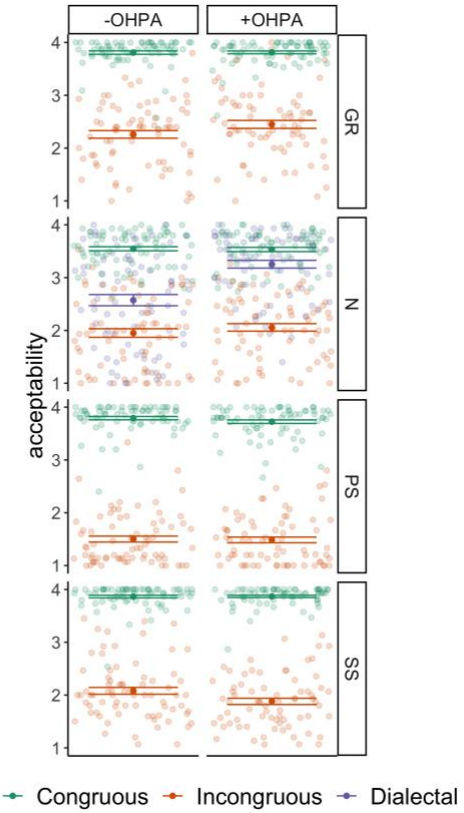


*Figure 5-1: Grand Average and By-subject Average Acceptability Judgment Results in -OHPA and +OHPA Groups.* The x-axis is a placeholder to show the in each condition: GR, N, PS, and SS. The preferences are congruous (green), incongruous (orange), and dialectal (purple). The dialectal preference is only used in the Needs condition. Acceptability is listed across the y-axis, where 1 is unacceptable and 4 is acceptable. The bolded dots represent the average acceptability of that preference in one condition for all participants within the group, and the standard error is shown by the error bars for each of the averages. The transparent dots represent the by-subject average for the preferences in each condition. [-OHPA (left; n=74) and +OHPA (right; n=73)].

In both groups and in all four conditions, the congruous preferences are rated more acceptable than the incongruous preferences. The by-subject average acceptability judgments help identify more categorical or more gradient conditions. The congruous preference in all four conditions shows the by-subject averages concentrated primarily between 3 and 4, yielding a categorical judgment as acceptable. The by-subject averages for the incongruous preferences for the Phrase Structure and Lexical Semantic conditions are primarily concentrated below 2, yielding a categorical unacceptable judgment. The incongruous preference for the Gender Reflexive and the Needs conditions shows the by-subject averages ranging across the entire Likert-Scale, but with most averages falling between 1-3. This range of judgments suggests the

incongruous preference in these conditions are more gradient in nature (see section 5.4 for a brief discussion on why the Gender Reflexive construction shows gradience in acceptability).

The key difference between the -OHPA and +OHPA groups lies in the dialectal preference. The -OHPA group shows gradient judgments with the by-subject average acceptability in all four points of the Likert-Scale and most of the averages fall between 1-3. The +OHPA group shows more categorical judgments with the by-subject averages falling between 3-4. This difference in the dialectal preference based on group membership suggests that the linguistic experience with a dialectal preference drives the difference in whether the group shows more categorical or gradient judgments.

### *5.3.2 Online Measure: RTs*

Figure 5-2 shows the averaged RTs in milliseconds from two words before and two words after the target word in all four conditions.

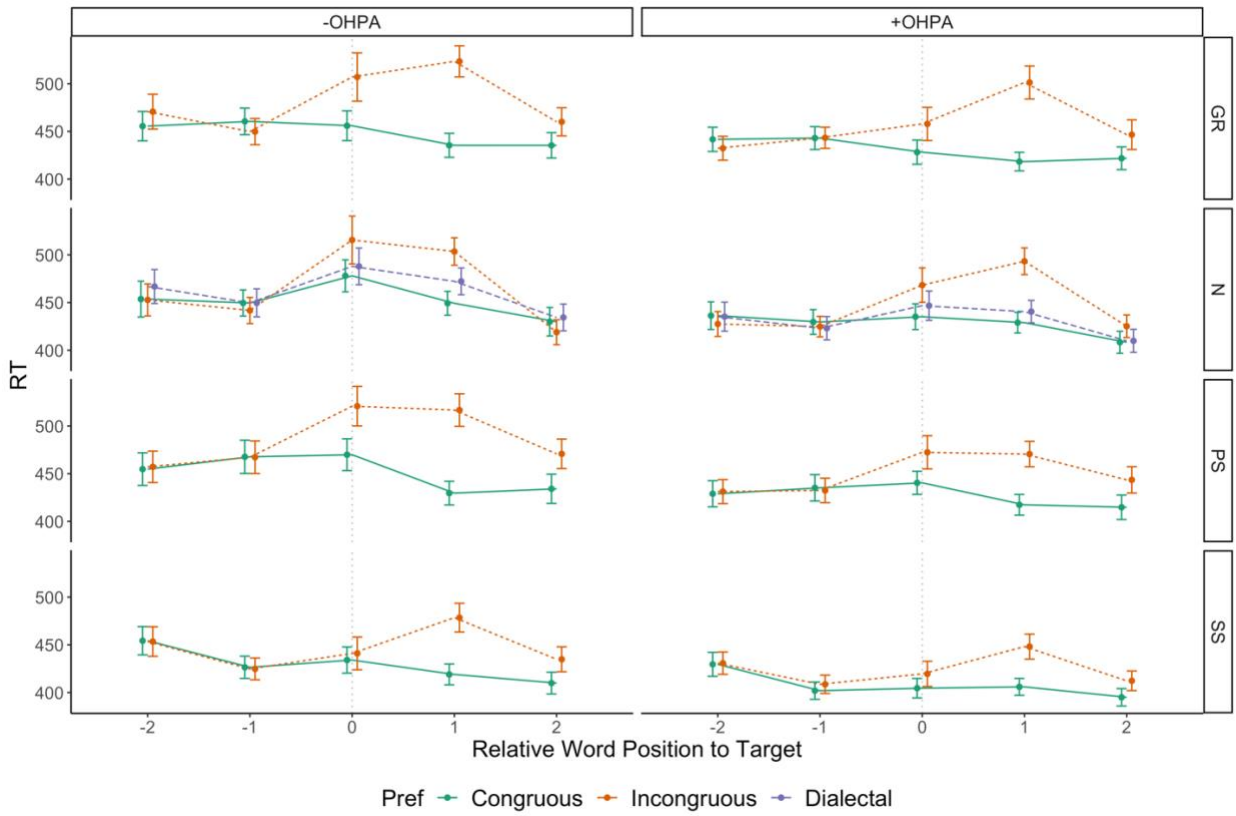


*Figure 5-2: Averaged Self-Paced Reading Time Results in -OHPA and +OHPA Groups.* The averaged RTs in milliseconds (y-axis) are shown by their position relative to the target word (x-axis), which is located at 0 and highlighted by the vertical grey dotted lines. Each horizontal row represents one of the four conditions in the experiment: Gender Reflexive (GR), Needs (N), Phrase Structure (PS), and Lexical Semantic Sentences (SS). The lines are colored green, orange, and purple to represent the preference of each condition: congruous, incongruous, and dialectal respectively. Only the Needs condition contains all three preferences. [-OHPA (left; n=74) and +OHPA (right; n=73)].

RTs are slower at Relative Word Position to the Target 0 and Target+1 in the incongruous preference (orange line) as compared to the congruous (green line) in the Gender Reflexive, Needs, and Phrase Structure conditions for both groups in Figure 5-2. For the lexical semantic condition in both groups, RTs are slower at Target+1 in the incongruous preference than in the congruous preference.

While both groups largely pattern the same, the dialectal preference (purple line) indicates RT differences between the groups. In the -OHPA group (left), the dialectal preference exhibits slower RTs than the congruous preference. However, the dialectal preference is read faster than the incongruous sentences. In the +OHPA group (right), the dialectal preference exhibits slightly slower RTs than the congruous preference, and faster RTs than the incongruous preference. This suggests that linguistic experience with the dialectal condition, regardless of personal usage, is driving the differences between the groups, and this suggests that linguistic experience modulates both online RTs and offline acceptability judgments.

### *5.3.3 Relationship of the Offline and Online Measures*

#### 5.3.3.1 Visual Relationship between Reading Time and Acceptability

Figure 5-3 combines the averaged self-paced RT and average acceptability judgment results across the four conditions. The y-axis on the average acceptability judgment graph has been transformed by -1 to better visualize the relationship between the average RT and average acceptability judgments.

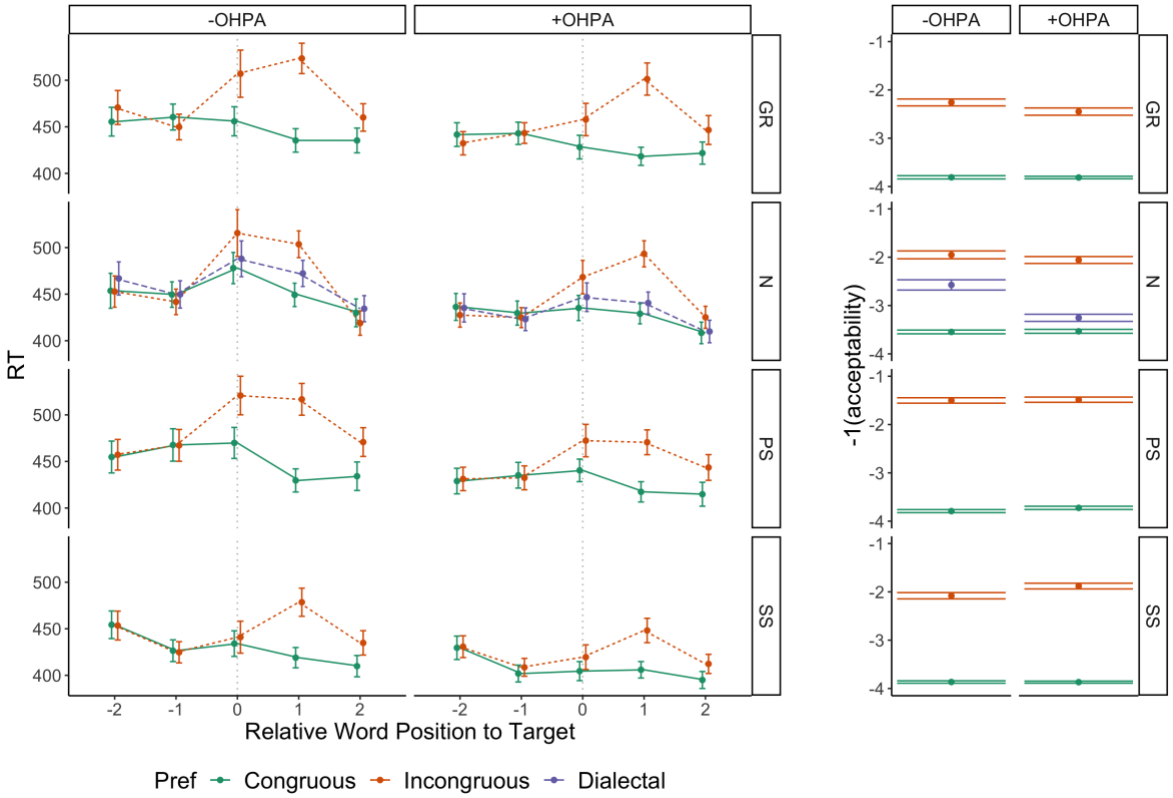


*Figure 5-3: Comparison of the Averaged Self-Paced Reading Time (left) and Averaged Acceptability Judgment (right) Results in -OHPA and +OHPA Groups.* Left: reprint of Figure *5-2*. Right: The averaged acceptability judgment responses (from

Figure *5-1*) have been multiplied by -1 to better visualize the relationship between RTs and acceptability. -1 is unacceptable and -4 is acceptable. The dots represent the average acceptability of that preference in one condition for all participants, and the standard error is shown by the error bars for each of the averages. [-OHPA (left; n=74) and +OHPA (right; n=73); Preferences: green= congruous, orange= incongruous, purple= dialectal; Conditions GR, N, PS, SS].

Slower RTs at Target+1 are reflected alongside lower acceptability ratings in the incongruous preferences across all four conditions across both groups. The baseline congruous

preferences contain the faster RTs at Target+1, as well as higher acceptability judgment ratings in both groups. The dialectal preference across the -OHPA and +OHPA groups show differences in both average RTs and average acceptability judgments. The average RT for the -OHPA group is faster at Target+1 than the incongruous preference, but the average RT tracks closer to the incongruous preference. Additionally, the average acceptability judgment is rated between 2 and 3, which is slightly more acceptable than the incongruous preference. In the +OHPA group, the average RT at Target+1 for the dialectal preference is faster than the incongruous preference, but the average RT tracks closer to the congruous baseline. The average acceptability for the dialectal preference is rated as higher than incongruous, but lower than congruous. This is likely due to language attitudes regarding the dialectal preference that the +OHPA group encounters (see section 6.2.2.3 for language attitudes participants express about needs+past participle).

In sum, both online RTs and offline acceptability judgments reflect dialect experience: +OHPA show higher judgments and faster RTs for dialectal preference, and -OHPA show lower judgments and slower RTs to the dialectal preference. Given this relationship, this leads then to the next question: is the degree of change in the online RTs proportional to the degree of change in the acceptability judgments? This will be examined through a linear mixed effects model in the following section, both with and without group as a fixed effect.

### 5.3.3.2 Results for the Linear Mixed Effects Model

Figure 5-4 demonstrates the relationship between acceptability and log-adjusted RTs for every trial for each participant across the four conditions.

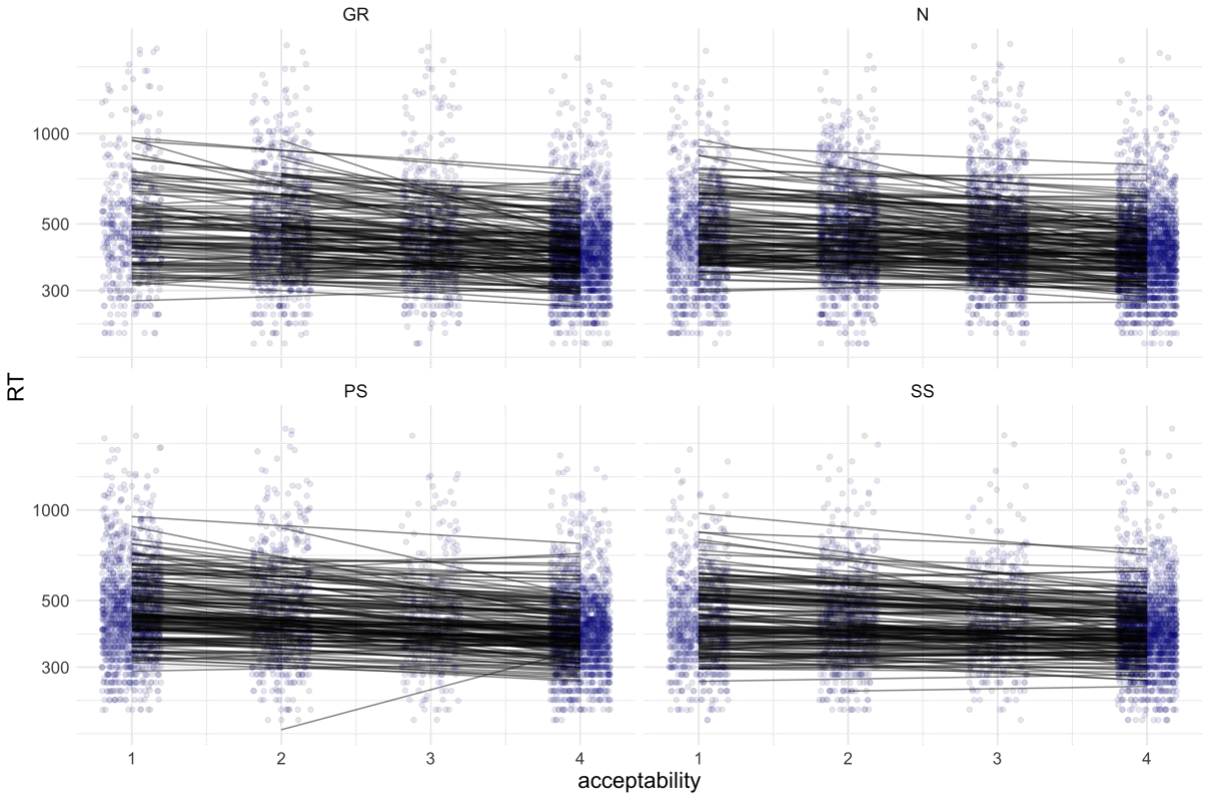


*Figure 5-4: Relationship between Acceptability and Reading Time.* Each panel represents one of the four conditions in the experiment: Gender Reflexive (GR), Needs (N), Phrase Structure (PS), and Lexical Semantic Sentences (SS). Each panel shows acceptability judgments across the x-axis and the log-adjusted RTs across the y-axis. The blue dots represent the acceptability judgments & RTs for every trial for each participant. The black lines represent the linear trend for each participant. [n=147].

Figure 5-4 shows that across all four conditions most participants show a slight negative relationship between the RTs and acceptability judgments. RTs for trials rated as unacceptable (1) tend to be higher than the RTs for trials rated as acceptable (4). Other participants show a flat or positive relationship between the RTs and acceptability judgments.

A linear mixed effects model was performed looking at the interaction between acceptability and condition as a function of RTs. The statistical summary of the linear trend in each condition is shown Table 5-5.

*Table 5-5: Statistical Summary of the Linear Trend per Condition.* The acceptability trend, lower HPD, and upper HPD are slopes across the four conditions in the experiment: Gender Reflexive (GR), Needs (N), Phrase Structure (PS), and Lexical Semantic Sentences (SS).

| Condition | Acceptability Trend | Lower HPD | Upper HPD |
|---|---|---|---|
| Gender Reflexive | -0.0643 | -0.0737 | -0.0544 |
| Needs | -0.0484 | -0.0574 | -0.0402 |
| Phrase Structure | -0.0515 | -0.0605 | -0.0434 |
| Lexical Semantic | -0.0330 | -0.0406 | -0.0254 |

This analysis suggests that a reliable negative trend (e.g., more acceptable sentences yield faster RTs) is observed for all four constructions. That is, the estimated linear relationship between acceptability and reading was distinguishable from zero, but it is slight. To determine if group impacts the relationship between acceptability judgments and RTs, group was added as a factor to the mode.

Table 5-6 shows the statistical summary of the linear trend per group and condition but focuses on the results of the dialectal variant.

*Table 5-6: Statistical Summary of the Linear Trend per Group and per Condition.* The acceptability trend, lower HPD, and upper HPD are slopes across the two groups (-OHPA and +OHPA) for the Needs condition.

| Group | Condition | Acceptability Trend | Lower HPD | Upper HPD |
|---|---|---|---|---|
| -OHPA | Needs | -0.0408 | -0.0522 | -0.0280 |
| +OHPA | Needs | -0.0568 | -0.0688 | -0.0447 |

When group is added into the model, the negative trend remains in for the *needs* conditions across both groups. This suggests that regardless of group membership, the relationship between RT and acceptability judgments is negatively correlated (e.g., as RTs decrease then acceptability judgments increase). Both groups show a proportional speedup in RT for higher-acceptably items. Again, although very slight, the estimated linear relationship between acceptability and reading was distinguishable from zero.

### 5.4 Discussion

The results of the experiment suggest that linguistic experience drives the difference in the online and offline measures for the dialectal preference in the *needs* condition, as Squires (2014a) had found with nonstandard *don't*. The predictions regarding the dialectal preference modulated by linguistic experience also behaved as expected. The RTs and acceptability judgments were modulated based on dialectal experience. Those who were expected to have dialectal experience had the offline and online measures patterned more like the congruous preference. The congruous preference showed faster RTs and higher acceptability in all four conditions. Those who weren't expected to have dialectal experience patterned more like the incongruous preference in these measures. The incongruous preference showed slower RTs and lower acceptability judgments.

This study shows a subtle difference in the offline acceptability judgment measures based on language experience. The -OHPA group (e.g., low language experience) showed gradience in acceptability with by-subject acceptability judgment averages in the dialectal needs construction distributed across all four-points of the Likert scale. In contrast, the +OHPA group (e.g., high language experience) showed the by-subject acceptability judgment averages in the dialectal needs construction distributed between 2-4 (although a couple by-subject averages were in 1).

The other results beyond the target dialectal condition behaved mostly as predicted. The incongruous preferences for the gender reflexive, phrase structure, and needs conditions show similar results to a syntactic violation across the two groups and this is expected as both groups have the same linguistic experience with the syntactic conditions. The incongruous preference in the lexical semantic condition shows similar results to a semantic violation across the two groups, which is expected as both groups are expected to have the same semantic knowledge.

Figure 5-1 showed gradience in acceptability for the incongruous gender reflexive condition, which was not predicted in Table 5-3. The by-subject acceptability judgments in both groups (and in Chapter 4) ranged across the entire four-point Likert scale, with a grand average acceptability of 2. However, this was not the case for the incongruous gender reflexive condition in Chapter 2. The difference in the acceptability judgment results for the incongruous gender reflexive condition may be a result of an increased social discussion regarding gender and pronouns in society that has increased over the past five years. The EEG study took place in fall 2019, while both self-paced reading studies took place in fall 2020 and fall 2021. The gradience in acceptability in the self-paced reading studies shows the perceptions of gender reflexive sentences changing, although the online RTs have not yet been impacted.

Both online RTs and offline acceptability judgments reflect dialect experience: +OHPA show higher judgments and faster RTs for dialectal preference, and -OHPA show lower judgments and slower RTs to the dialectal preference. This leads to both groups showing negative correlations between acceptability judgment and RTs in the dialectal preference, as well as all other conditions. These results in this study align with the results in Chapter 4 in that the correlation between acceptability judgments and RTs is slightly. This current study showed a replication for the gender reflexive and lexical semantic conditions, and new contributions for the phrase structure and needs conditions.

Language experience modulates the processing of the dialectal preference as expected for the online RTs and offline acceptability judgments, which differs from the results from García (2017) and Zaharchuk et al. (2021). García (2017) found a mismatch between the online ERP measures and the acceptability judgments in the bidialectal speaker group, in that no P600 response (expected) and lower than predicted acceptability judgment to the omitted -S condition

(unexpected). In contrast, Zaharchuk et al. (2021) found a mismatch between the online ERP measures and the acceptability judgments in the familiar group. A P600 response was elicited to the attested double modal condition (unexpected) and the acceptability of the attested double modal condition was rated acceptable (expected).

Differences in experimental protocols likely explain the conflicting results between the current study and the studies from García (2017) and Zaharchuk et al. (2021). The first difference was the two different online measures, RTs and P600 responses, that were collected in-person and via the internet. Perhaps the relationship between the online and offline measures is restricted to RTs and not ERP amplitudes. The second difference was in the ordering of the experimental tasks. The current study collected the online and offline measures to every trial, which was used to capture the participants immediate ratings to each sentence. The other two studies conducted a post-EEG experiment acceptability judgment task, which is a cognitively different task for participants and one that is also untimed. The untimed norming data in section 2.3.1 showed categorical acceptability judgment responses to the subject-verb agreement condition; however, gradient responses were found to the same condition when acceptability was collected immediately after stimulus presentation. The final difference was the modality of the stimuli, but several studies have shown that the P600 response to morphosyntactic violations are elicited in both the auditory and visual modality without significant differences in latency (Balconi & Pozzoli, 2005; Hagoort & Brown, 2000; Hansen, 2005). A future study should be modified to address these differences and reconcile the conflicting results between the SPR and ERP studies (see section 6.2.1 for how the study could be modified to address these differences).

## 5.5 Summary of the Study

This study probes whether language exposure to a dialectal condition modulates the negative relationship between acceptability judgments and RTs established in Chapter 4. The needs+past participle condition is likely to exhibit gradience in the acceptability judgment responses based on location of the participants. Therefore, the participants were recruited in two groups based on language exposure based on current location and location born. These two groups read and immediately rated the acceptability of each sentence across four conditions: needs, phrase structure, gender reflexive, and lexical semantic. A linear mixed effects model was fit with RT as a dependent measure and acceptability, condition, group, and their interactions as a fixed effect. The model included subjects as a random intercept and random slopes for their acceptability, condition, and their interaction. The results show a small, but statistically reliable negative relationship between RTs and acceptability judgments across all four conditions and both groups, in that as acceptability judgments increase, RTs decrease. Overall, regardless of language exposure, a proportional relationship between offline and online measures was found within the dialectal condition suggesting that language exposure affects offline and online measures similarly when collected within a single trial.

# Chapter 6 Conclusion

This dissertation developed a methodology to collect online and offline measures within the same trial and to analyze whether gradience in offline acceptability judgments show proportional gradience in online measures. The results from this procedure were affected by the experimental methodology, such that the SPR methodology showed a proportional relationship between the online and offline measures. I conclude with a summary of contributions in section 6.1 and directions for future work in section 6.2 with a primary focus on the impact of sociolinguistic awareness on the relationship between online and offline measures.

## 6.1 Summary of Contributions

This dissertation contributes to our understanding of how gradience within a syntactic construction, the type of online measure, and language experience impacts the relationship between online and offline measures. Overall, the results of the three experiments suggest a slight negative correlation between the online and offline measures in the two SPR experiments, but not in EEG. The EEG experiment (Chapter 2) only showed a reliable negative correlation for the subject-verb agreement construction, while the two SPR experiments (Chapter 4 and Chapter 5) showed a reliable negative correlation for all constructions. As such, a proportional relationship between online and offline measures is more reliably represented for RTs as the online measure than ERP amplitudes. Given this, I conclude that proportional gradience between online and offline measures is an outcome for the SPR experimental methodology and online RTS, but there is not enough evidence to show the same outcome for the EEG methodology and

online neural signals (see section 6.2.1 for a discussion on rerunning the EEG experiment with SPR protocol).

Chapter 1 examined non-syntactic cognitive factors that affects acceptability judgments to see whether these factors similarly affect online measures. Chapter 5 showed that language experience based on geographical location did not modulate the negative correlation between online and offline measures in the dialectal construction using SPR via internet collection. Previous research suggested that language users may report dialectal variants as unacceptable and yet still unconsciously use these variants (García, 2017; Hildebrandt, 2017; Labov, 1996). The acceptability judgments, however, aligned with the prediction that those with language experience would rate the dialectal condition on average higher than those who lacked the language experience. Individual differences within both groups were accounted for through the correlation between the online and offline measures, and any variability in the offline acceptability judgments likely was evident in the online RTs.

One area that was not explored in a detailed analysis was syntactic satiation as a task effect (as discussed in section 1.2.2). Hildebrandt and Brennan (2022) used the RTs and acceptability judgment data from Chapter 5 to test whether the acceptability effects in showed satiation effects across all four conditions (e.g., Phrase Structure, Needs, Gender Reflexive, and Lexical Semantic). Performance constraints typically are susceptible to satiation, but not effects reflecting grammatical constraints (Sprouse, 2022). All four constructions showed adaptation effects in the RTs throughout the five blocks (e.g., RTs became faster in each block of the experiment). However, the acceptability judgments remained stable throughout the five blocks of the experiment in both the +OHPA and -OHPA groups. This stability of the acceptability judgments suggests that these judgements reflect a grammatical constraint which are impervious

to satiation effects. Overall, the results suggest that the offline measure exhibits stability, while the online measure adapts over the course of the experiment. But crucially, this difference in task effect (e.g., satiation versus adaptation) did not affect the negative correlation between the online and offline measures in Chapter 5. As sources of gradience in the offline acceptability judgments, task effects and language experience in Chapter 5 do not modulate the proportional relationship between the online and offline measures.

In fact, this dissertation found that gradience in acceptability within a variable syntactic construction affects online measures similarly. The variable realized complementizer condition in the EEG experiment (Chapter 2) showed a trend toward a P600 response but was not statistically significant. The same condition in the SPR experiment (Chapter 4) showed slower RTs in sentences without the complementizer, but these RTs were not reliably different than sentences with the complementizer. The acceptability judgment responses to the subject-verb agreement violations were gradient in both the EEG and SPR experiments. The subject-verb agreement condition in the EEG experiment also showed a trend toward a P600 response but was not statistically significant, and the RTs in the SPR experiment showed reliably slower RTs in subject-verb disagreements. This difference between statistical significance between the EEG and SPR experiments should be taken cautiously, as even the gender reflexive filler condition was not statistically significant in the EEG experiment.

Additionally, this dissertation provided evidence that syntactic preference (Chapter 2 and Chapter 4) and language experience (Chapter 5) modulate gradience similarly in the online and offline domains. For syntactic preference, the complementizer condition may have shown an attenuated P600 response, in line with Osterhout et al. (1994) as two syntactic options were possible. As discussed above, each online measure did not show reliable differences, but trended

towards reliability. The acceptability judgments in Chapter 2 and Chapter 4 both showed reliable differences; however, these differences were small as the acceptability judgment averages were similar (see Figure 2-3 and Figure 4-2). The language experience of the participants also modulated the online and offline measures for the gradient dialectal variant. For those with experience, RTs were faster and acceptability judgments were higher and those without experience had slower RTs and lower acceptability judgments.

## 6.2 Directions for Future Work

This dissertation pushes forward on the conversation about whether the relationship between online and offline measure is proportional. Additional work is needed to address the difference found between the online measures to determine if ERP amplitudes are reliable measures to assess the relationship between online and offline measures (section 6.2.1). Chapter 5 concluded that language experience did not modulate the negative correlation between RTs and acceptability judgments. A next step would be to assess how awareness and indexing impact this relationship. Preliminary visualizations are presented in section 6.2.2, which suggest that awareness and indexing groupings should uphold the negative trending relationship.

### *6.2.1 Returning to EEG*

The nonsignificance of the P600 responses from Chapter 2 should be examined more closely due to the potential noise in the data. First of all, data collection stopped short of my goal of 30 participants due to hardware issues with the EEG. These issues likely caused noise within the set of data I was able to collect, and it limited increasing the number of participants. Secondly, I collected data from 29 participants and 22 participants data were able to be used in the analysis. A larger group of participants would be beneficial to reduce the noise within the

EEG data and this may change the reliability of the online ERP amplitude measure. Lastly, the RSVP reading protocol may have added noise to the EEG data, as participants may not have been actively engaged with the task. Recalling this discussion from section 5.4, changing the reading protocol to SPR shouldn't affect the online results.

Considering these factors, I suggest increasing the number of participants to 60 and rerun the experiment in Chapter 2 with the same constructions. Additionally, I would change the reading protocol to SPR to collect two online measures, ERP amplitudes and RTs, to gain a better sense of whether the relationship between acceptability and these online measures is proportional within a single trial in the same experiment. Then, if ERP amplitude and acceptability show a negative correlation, I would replicate the experiment in Chapter 5 using EEG with the SPR reading protocol in order to provide evidence against or supporting García (2017) and Zaharchuk et al. (2021) but the study would differ in the single trial data collection.

### *6.2.2 Exploratory Visualizations: Awareness and Indexing of Dialectal Needs*

This visual analysis explores the effect of the online and offline measures from Chapter 5 when grouped by the awareness and the indexation of the dialectal preference. As awareness has many definitions within the sociolinguistic literature, I am adopting Squires' (2016) definition of awareness as perceiving and explicitly noticing, which were gathered from answers to a post-experiment sociolinguistic survey (Appendix G). These answers were coded and used to determine new groupings based on the awareness of self, awareness of others and the indexation of the dialectal *needs* condition. These awareness and indexation questions were optional, which left n=146 for the awareness groupings and n=102 for the indexation groupings.

Participants responded yes or no if they themselves use the dialectal *needs* construction (Awareness of Self) and they responded yes or no if others they know use the dialectal *needs*

construction (Awareness of Others). Lastly, participants provided answers about who do they think uses the dialectal *needs* construction. Answers were sorted into similar categories and the only the top three indexing themes are included in the results. The top three indexing themes are: Anyone (n=33), English Language Learners (n=51), and the Working Population (n=18)[33].

The following results sections show the online RTs and offline acceptability judgments for awareness of self, awareness of others, and the top indexation themes in the *needs* condition. See Appendices I, J, and K to observe the complete results of the awareness and indexation groupings across all four conditions.

[33] Although +/- OHPA groupings aren't analyzed here, it is interesting to note that only 2 participants from -OHPA indicated that the working population would use the dialectal *needs* preference.

### 6.2.2.1 The Effect of Awareness of Self

Figure 6-1 (top) shows the mean RTs in milliseconds at the target word +1 and Figure 6-1 (bottom) shows average acceptability judgments in the *needs* condition. Participants self-reported if they themselves use the dialectal preference (bottom right panel) or do not use the dialectal preference (bottom left panel).

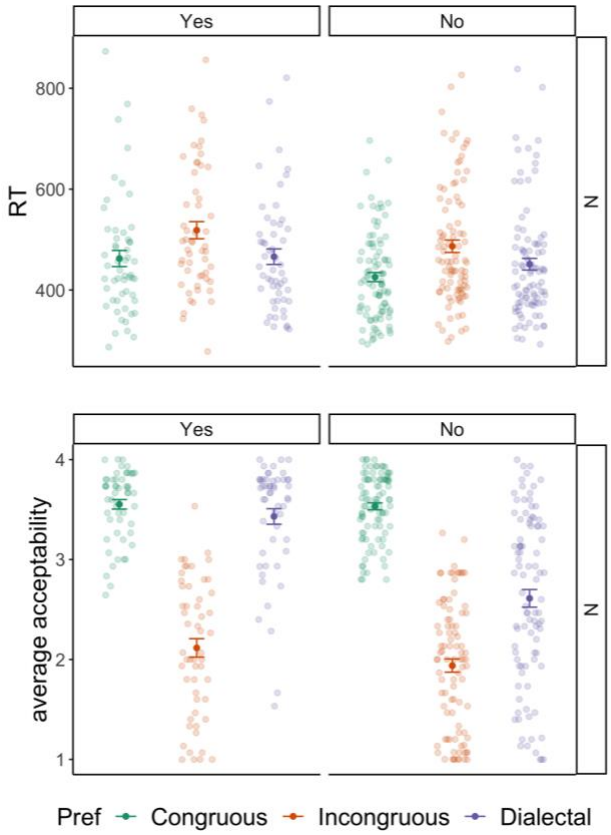


*Figure 6-1: Average Reading Time (top) and Average Acceptability Judgments (bottom) Grouped by Awareness of Self in the Dialectal Needs Condition*. Top: The x-axis is all the preferences and RT is listed across the y-axis in milliseconds. Bottom: The x-axis is all the preferences. Acceptability is listed across the y-axis, where 1 is *unacceptable* and 4 is *acceptable*. Participants answered if they used needs sentences with yes (left) or no (right). The preference of each condition are lines are congruous (green), incongruous (orange), and dialectal (purple). The bolded dots represent the average RT or acceptability of that preference in the *needs* condition for all participants within the group, and the standard error is shown by the error bars for each of the averages. The transparent dots represent the by-subject RT or acceptability average for the preferences in the *needs* condition. [Yes n=53; No n=93].

The incongruous preference is read more slowly than the other two preferences across both groups. The dialectal and congruous preferences exhibit different behavior based on self-

awareness of the participants. Those who use the dialectal preference have near-identical average RTs to the congruous preference with significant overlap in the standard errors of the mean RT. Those who do not use the dialectal preference shows a stacking relationship of the mean RTs in that the incongruous preference has the slowest RTs, followed by the dialectal, and then by the congruous preference that has the fastest RTs.

Across both groups, the incongruous and congruous preferences are rated similarly. The average acceptability judgment for incongruous preference is rated the lowest, while the average acceptability judgment for the congruous preference is rated the highest. The dialectal preference differs based on if the participants themselves use the dialectal preference. The yes group shows an average acceptability judgment close to the congruous preference, which the standard errors of the mean acceptability slightly overlap. The no group shows an average acceptability judgment between the congruous and incongruous preferences (around 2.5).

#### 6.2.2.1 The Effect of Awareness of Others

Similar results are found when participants are group by if they are aware if other individuals that they know use the dialectal *needs* preference. The RT results are visualized in Figure 6-2 (top) and the acceptability judgment results are visualized in  Figure 6-2 (bottom).

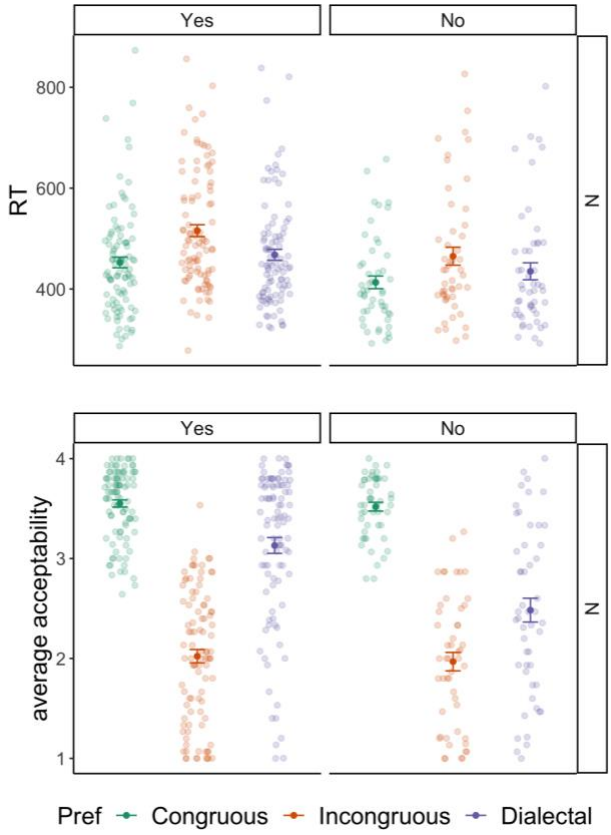


*Figure 6-2: Average Reading Time (top) and Average Acceptability Judgments (bottom) Grouped by Awareness of Others in the Needs Condition.* Top: The x-axis is all the preferences and RT is listed across the y-axis in milliseconds. Bottom: The x-axis is all the preferences. Acceptability is listed across the y-axis, where 1 is *unacceptable* and 4 is *acceptable*. Participants answered if others they know used *needs* sentences with yes (left) or no (right). The preference of each condition are lines are congruous (green), incongruous (orange), and dialectal (purple). The bolded dots represent the average RT or acceptability of that preference in the *needs* condition for all participants within the group, and the standard error is shown by the error bars for each of the averages. The transparent dots represent the by-subject RT or acceptability average for the preferences in the *needs* condition. [Yes n=96; No n=50].

The incongruous preference is read the slowest across both groups, while the congruous preference is read the fastest across both groups. The yes group (left panel) shows the average RT in the dialectal preference closer to the average RT in the congruous preference. The no group (right panel) shows the mean RTs ordering as the incongruous preference has the slowest

RTs, followed by the dialectal, and then by the congruous preference that has the fastest RTs. However, the dialectal preference is not statistically different than the congruous and incongruous preferences.

When comparing the average acceptability judgments across groups in, the congruous preference is rated as the most acceptable and the incongruous preference is rated as the least acceptable. In the dialectal preference, if the participants know someone else who uses the dialectal preference, the average acceptability is rated closer to the congruous preference. If the participants don't know someone else who uses the dialectal preference, then the average acceptability is rated closer to the incongruous preference.

### 6.2.2.2 The Effect of Indexical Themes

Figure 6-3 (top) shows the mean RTs in milliseconds at the target word +1 and Figure 6-3 (bottom) shows the average acceptability in the *needs* condition grouped by who might say the dialectal preference. The top three answers were anyone (left), English language learners (middle), and the working population (right).

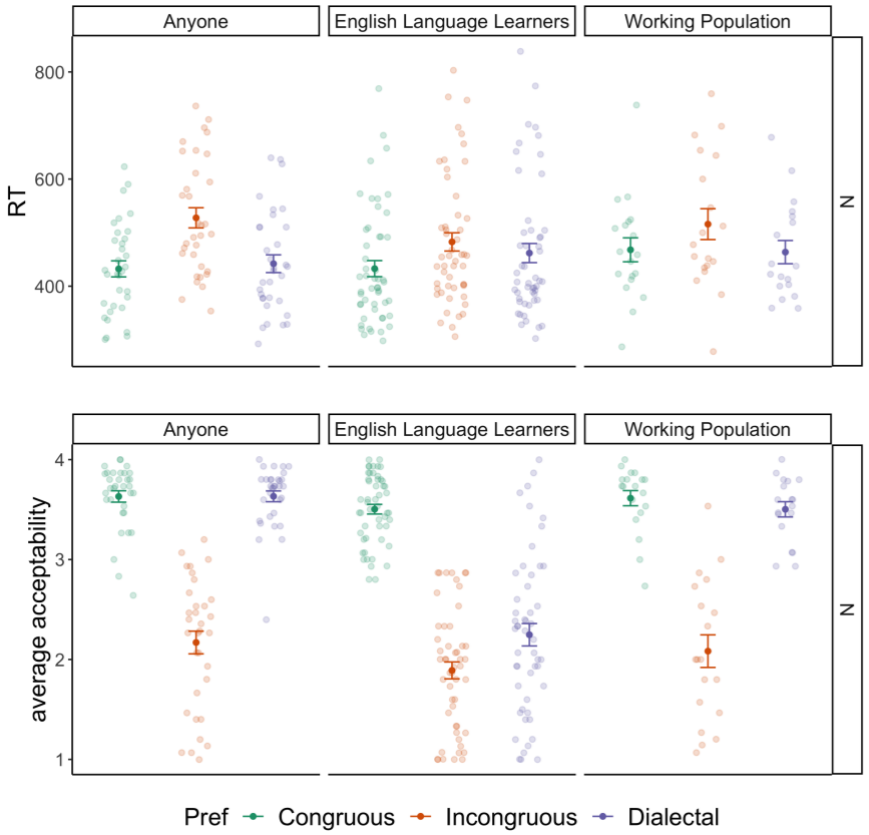


*Figure 6-3: Average Reading Time (top) and Average Acceptability Judgments (bottom) Grouped by Indexing in the Needs Condition.* Top: The x-axis is all the preferences and RT is listed across the y-axis in milliseconds. Bottom: The x-axis is all the preferences. Acceptability is listed across the y-axis, where 1 is *unacceptable* and 4 is *acceptable*. Participants answered who they think would use *needs* sentences and the top three themes are listed on the top of the graph Anyone (left), English Language Learners (middle), and Working Population (right). The preference of each condition are lines are congruous (green), incongruous (orange), and dialectal (purple). The bolded dots represent the average RT or acceptability of that preference in the *needs* condition for all participants within the group, and the standard error is shown by the error bars for each of the averages. The transparent dots represent the by-subject RT or acceptability average for the preferences in the *needs* condition. [Anyone n=33, English Language Learners n=51, Working Population n=18].

The average RT for the incongruous preference is read the slowest of across all three preferences and indexing themes. In the anyone and working population group, the incongruous

preference is read more slowly than the other two preferences. The dialectal and congruous preferences have near-identical average RTs with significant overlap in the standard errors of the mean RT. The English language learners grouping shows significant overlap in the standard errors of the average RT across all three preferences. The incongruous preference has the slowest RTs, followed by the dialectal preference RTs, and the congruous preference RTs are the fastest.

Across all indexing groupings the incongruous preference is rated the lowest on average as unacceptable (around 2). The average acceptability judgment for the congruous preference is rated as acceptable (above 3.5). One key difference for the average acceptability of the congruous preference is in the English language learners (ELL) group is that this average is lower than the other two indexing groups. The lower average acceptability judgment rating is also shown for the dialectal preference in the ELL group, with the dialectal preference being closer to the incongruous preference. In the anyone grouping (bottom left panel), the dialectal and congruous preference have near-identical acceptability judgment averages with significant overlap in the standard errors of mean acceptability. In the working population grouping (bottom right panel), the congruous preference acceptability judgment average is slightly higher than the dialectal preference, but the difference in preferences is not statistically reliable. Overall, it appears that language attitudes associated with English language learners impacted how acceptable the participants thought of the dialectal condition, but not how the other three conditions were rated (see Appendix 6-C).

#### 6.2.2.3 Discussions of the Preliminary Awareness and Indexation Trends

The results above focus on grouping the participants by their answers to the sociolinguistic survey questions about awareness and indexation regarding dialectal *needs* variant. However, participants showed awareness of other constructions when they answered a

question about what they noticed about the experiment. Out of the 147 participants, the most noticed condition was the gender reflexive condition in which 55% of the participants commented on this construction (n=81). Following that, 24% of the participants commented on dialectal needs (n=35), 16% commented on the phrase structure violations (n=24), and 8% commented on the lexical semantic stimuli (n=12). This suggests that the *needs* construction was noticeable, but to a lesser degree than the gender reflexive sentences.

One limitation is in how much the participants noticed the experimental constructions because the participants were not required to answer question 7 ("What did you notice about the experiment? Did anything stand out to you while you read and rated the sentences?") in the post-experiment survey. This led to a wide variety of answers in which some participants identified multiple constructions, others commented only one, and some instead commented on the experimental task. For future research, it would be advantageous to ask participants what they noticed after selected blocks in the experiment to specifically see what they are perceiving and explicitly noticing over the course of the experiment.

A few of the answers to the "what did you notice" questions had speakers who use dialectal *need* variant and their responses showed that they understand the construction linguistically and socially (e.g., they recognize the language attitudes associated with the variant). In writing about the sentences with dialectal *needs* one speaker in their 50s from the OPHA group said, "Many others sounded correct to me because of my local dialect but I know they are not technically considered correct." Another participant in their 20s from the OHPA groups said, "Yes, I saw multiple reoccurrences linguistically. Coming out of Appalachia myself, I was born and raised saying the (needs + past participle) formulation. I found it really interesting you guys are using that as an example here in this survey. Most people from America do not

even know that exists." Both participants recognize that they use the construction themselves, but they know that others either do not have knowledge of the construction or others deem it incorrect.

The visual results for awareness of self and others mirrors the patterns that were found for the +/-OHPA groupings. Answers of *yes* align with the +OHPA group (e.g., faster RTs and higher acceptability judgments for the dialectal preference) and answers of *no* align with the -OHPA (e.g., slower RTs and lower acceptability judgments for the dialectal preference)). Given this result, the relationship between these measures when grouped by awareness of self or awareness of others should maintain a negative correlation between the online RTs and offline acceptability judgment measures.

The Indexation themes are more complicated and might show differences in the proportional relationship between the online and offline measures. Specifically in the case of the English Language Learners category, the offline acceptability judgments were significantly lower, and the RTs were not reliably different in the dialectal preference. The lower acceptability judgment responses seem to follow Squires (2014b), where she found that participants assigned high status speakers to the standard construction and low status speakers to the nonstandard construction in a mouse click task. Fifty-one participants (35%) in this post-experiment question on indexing used terms related to English Language Learners (ELL) for dialectal *needs.* When we examine the indexation based on geographical location, 25.5% of the participants from the +OHPA group (n=13) and 74.5% of the participants from the -OHPA (n=38) used terms related to ELL in answering the indexation question. This ELL grouping difference suggests that ideas about who might use a specific construction, even if not directly asked about it until post-experiment, may have implications on the results.

### *6.2.3 Summary of Future Directions of Research*

These directions for future research would provide further insight on whether the relationship between online and offline measure is proportional. Specifically, future work would clarify whether ERP amplitudes as an online measure could be used to assess the relationship to offline acceptability judgments or whether EEG signals are too noisy, which obscures establishing a reliable relationship. With respect to sociolinguistic awareness, future work can solidify how awareness and indexation modulate the relationship between RTs and acceptability judgments through an experiment that specifically targets awareness and indexation.

## Appendix A: Individual Differences in Acceptability Judgments

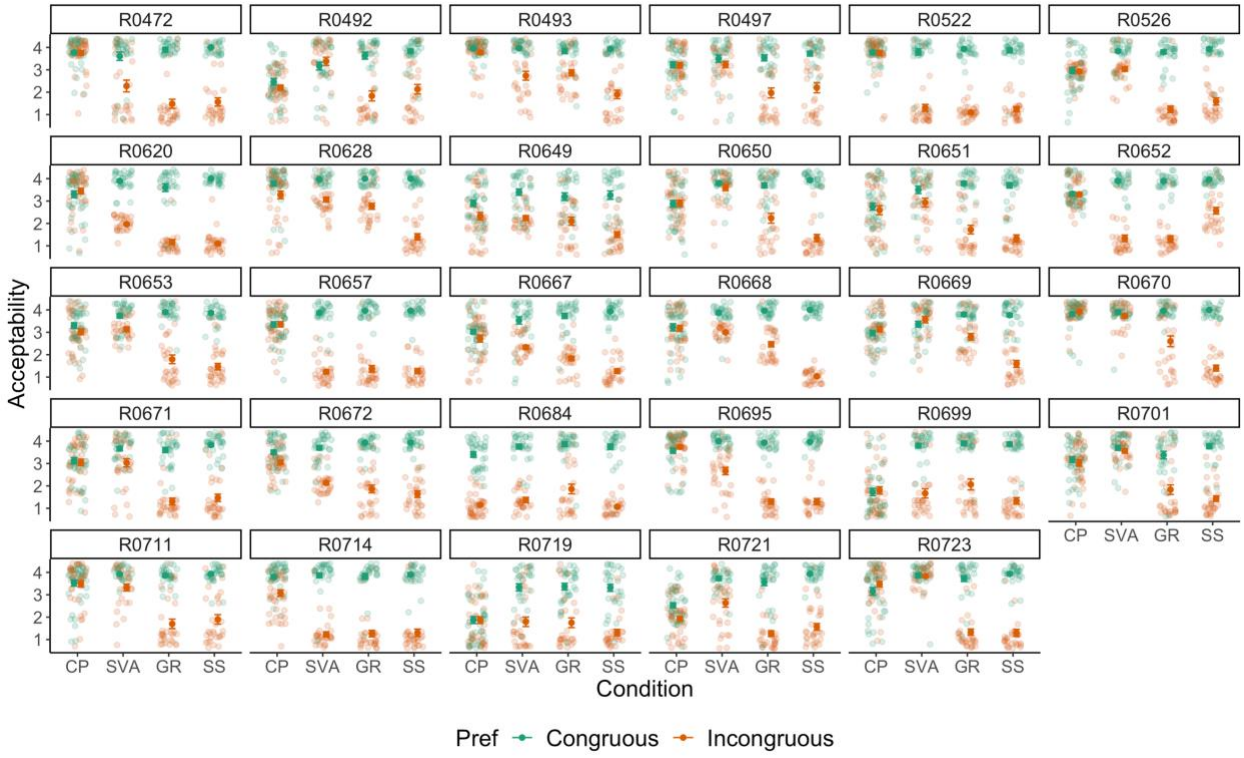


*Figure A-1: Grand Average and Token Acceptability Judgments across Subjects from the EEG Experiment. [n=29].* Condition is listed across the x-axis and contains all four conditions: Complementizer (CP), Subject-Verb Agreement (SVA), Gender Reflexive (GR), and Lexical Semantic (SS). Acceptability from 1 to 4 is listed across the y-axis, where 1 is unacceptable and 4 is acceptable. Each panel shows an individual subject's token acceptability judgments as well as a grand average acceptability with standard error by condition and preference.

Categorical ratings are shown as splits between the congruous (green) and the incongruous (orange) acceptability judgments. These are most evident in R0522, R0657, R0714 in the SVA, GR, and SS conditions. Gradient ratings show up as a mixture between the congruous (green) and the incongruous (orange) acceptability judgments. These are most evident in R0649, R0650, and R0721 in the CP condition, where we see the congruous and incongruous acceptability judgments across the entire rating scale.

## Appendix B: ERP Images of Congruous and Incongruous Stimuli

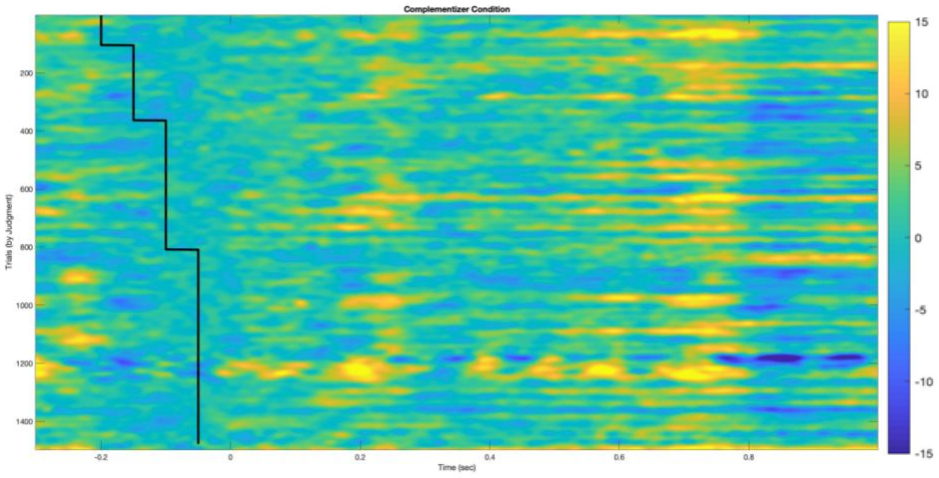


*Figure B-1: Complete Complementizer Condition Results.* Time is represented along the x-axis, while the P600 response is measures in voltage and shown in color. The color scale, which goes from yellow (high positivity; 15), orange (medium positivity), green (low positivity), lighter blue (low negativity), dark blue (high negativity; -15). Each row in the ERP-image is organized by the given acceptability response for each trial (black line from unacceptable (top) to acceptable (bottom)) on the y-axis. The area of interest for the Complementizer condition is between 0.6-0.8 seconds. [Electrodes: Pz, CPz, Cz, CP3, CP1, P1. n=22.]

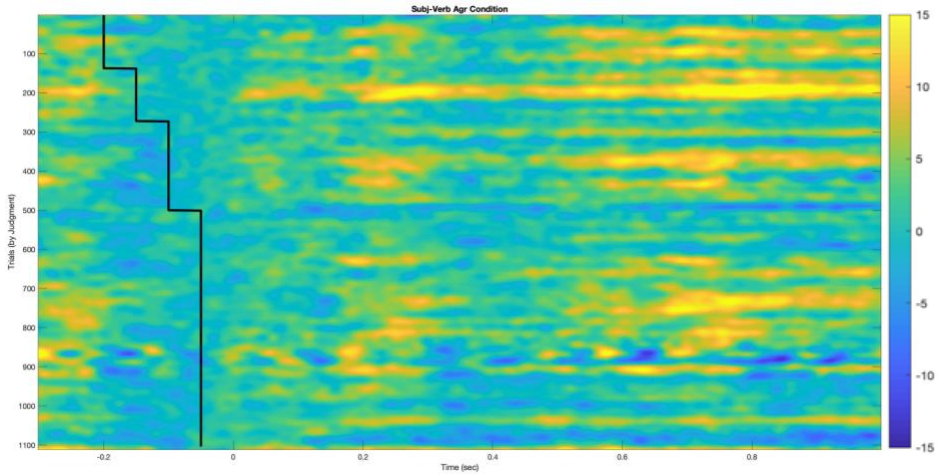


*Figure B-2: Complete Subject Verb Agreement Condition Results.* The area of interest for the Subject-Verb Agreement condition is between 0.6-1 seconds. [Electrodes: Pz, CPz, Cz, CP3, CP1, P1. n=22.]

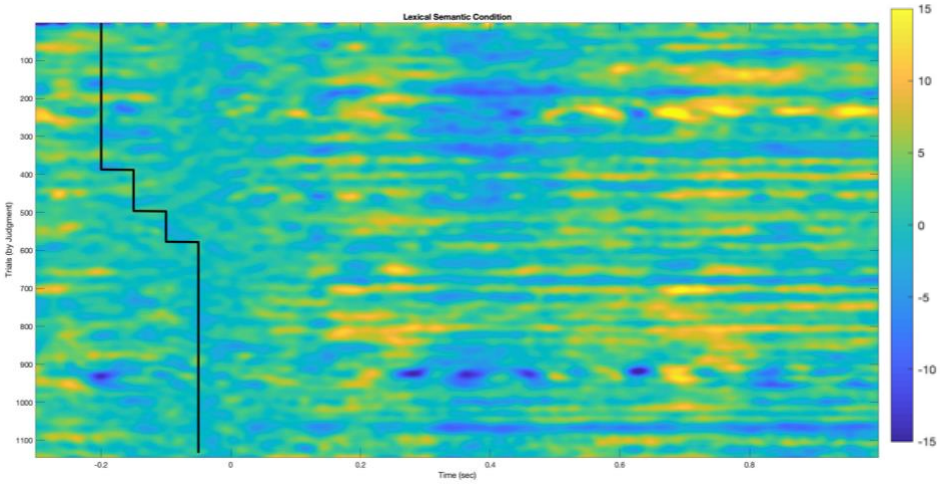


*Figure B-3: Complete Lexical Semantic Condition Results.* The area of interest for the Lexical Semantic condition is between 0.3-0.5 seconds. [Electrodes: Pz, CPz, Cz, CP3, CP1, P1. n=22.]

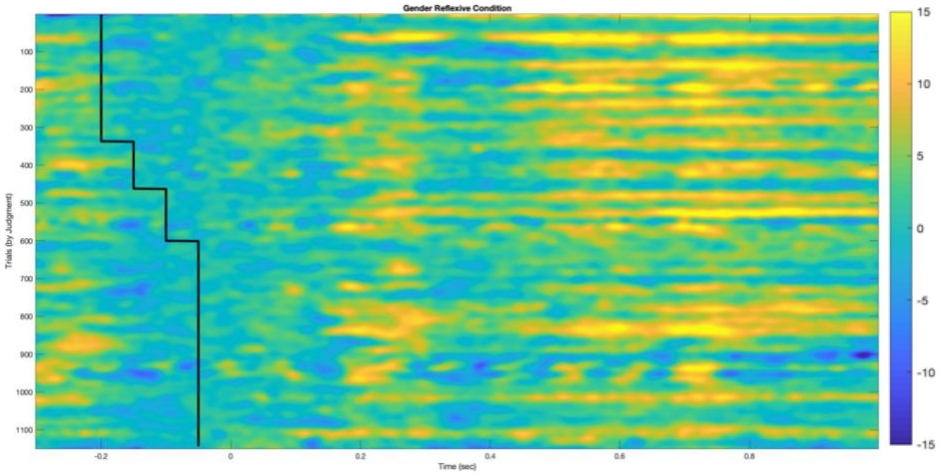


*Figure B-4: Complete Gender Reflexive Condition Results.* The area of interest for the Gender Reflexive condition is between 0.4-1 seconds. [Electrodes: Pz, CPz, Cz, CP3, CP1, P1. n=22].

## Appendix C: By-Subject Trend for the Amplitude and Acceptability Relationship

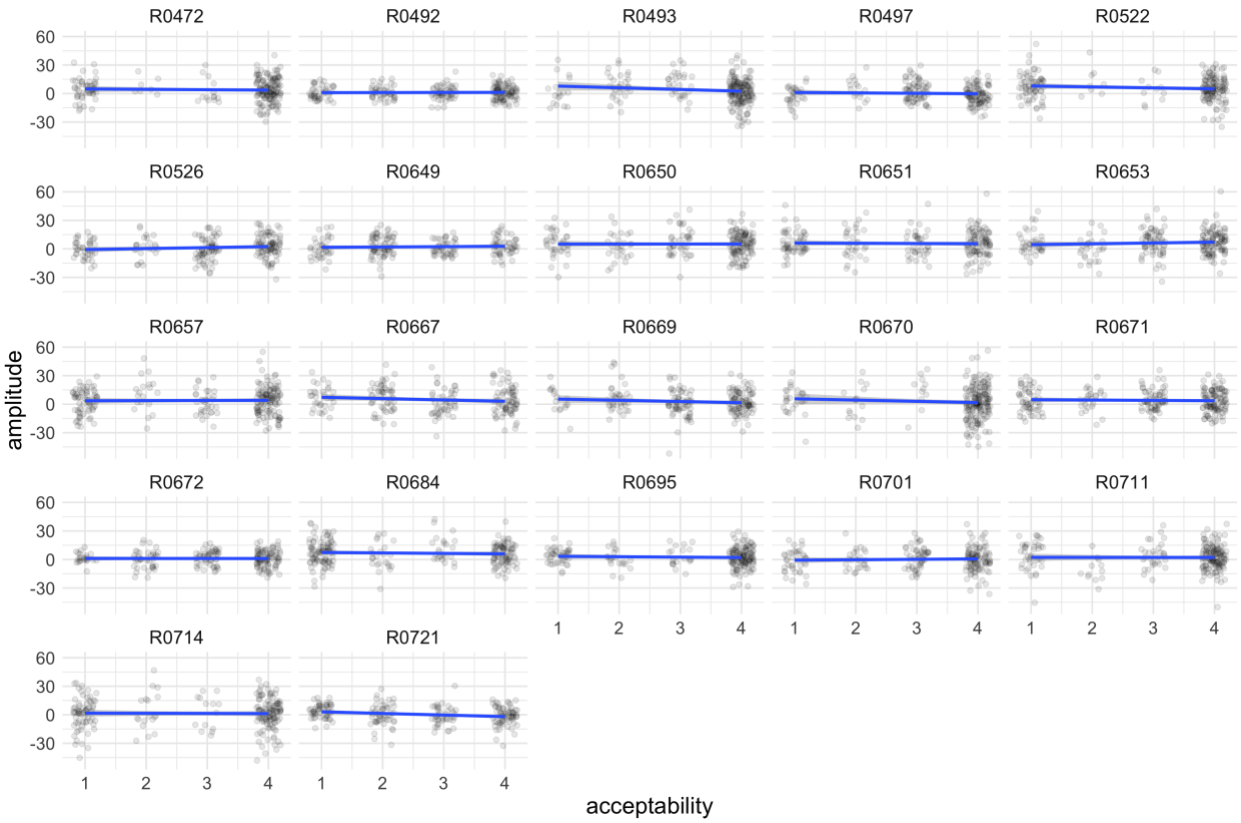


*Figure C-1: By-Subject Trend between Acceptability and Amplitude Relationship.* Each panel shows acceptability judgments across the x-axis and amplitude in microvolts across the y-axis. The black dots represent the acceptability judgments & RTs for every trial for each participant. The blue lines represent the linear trend within participant. Most participant acceptability judgment trends appear to be flat or slightly negative. [Electrodes: Pz, CPz, Cz, CP3, CP1, P1. n=22.]

## Appendix D: Practice Block Sentences

*Table D-1: Chapter 4 Practice Block Sentences.*

| Condition | Type | Sentence |
|---|---|---|
| Adverb | Congruous | We could quickly cook some dinner. |
| Adverb | Dialectal | We could quick cook some dinner. |
| Adverb | Incongruous | We could quicker cook some dinner. |
| Lexical Semantic | Congruous | My neighbor witnessed the murder last night in the alley |
| Lexical Semantic | Congruous | Patty mowed the lawn yesterday morning after breakfast. |
| Lexical Semantic | Incongruous | During the expedition, the archaeologists explored a cocktail to look for Egyptian artifacts. |
| Lexical Semantic | Incongruous | Kelley went to the bar and ordered a cave for herself. |

## Appendix E: Sociolinguistic Survey

1. What is your age in years?

2. How do you currently describe your gender identity?

3. What are languages that you speak and level of proficiency?

4. In what state were you born?

5. Approximately how long did you stay in your birth state?

6. Where have you lived previously and for how long?

7. What did you notice about the experiment? Did anything stand out to you while you read and rated the sentences?

## Appendix F: Complete List of Practice Block Sentences

*Table F-1: Chapter 5 Practice Block Sentences.*

| Condition | Type | Sentence |
|---|---|---|
| Adverb | Congruous | We could quickly cook some dinner. |
| Adverb | Dialectal | We could quick cook some dinner. |
| Adverb | Incongruous | We could quicker cook some dinner. |
| Adverb | Congruous | My brother slowly walked to the store in the rain. |
| Adverb | Dialectal | My brother slow walked to the store in the rain. |
| Adverb | Incongruous | My brother slower walked to the store in the rain. |
| Adverb | Congruous | My roommate quickly wrote an email to their professor. |
| Adverb | Dialectal | My roommate quick wrote an email to their professor. |
| Adverb | Incongruous | My roommate quicker wrote an email to their professor. |
| Adverb | Congruous | He will quickly make the bed. |
| Adverb | Dialectal | He will quick make the bed. |
| Adverb | Incongruous | He will quicker make the bed. |
| | | |
| Phrase Structure | Congruous | I drank Lisa's brandy by the fire. |
| Phrase Structure | Incongruous | I drank Lisa's by brandy the fire. |
| Phrase Structure | Congruous | She sold Raymond's gold on his birthday. |
| Phrase Structure | Incongruous | She sold Raymond's on gold his birthday. |
| Phrase Structure | Congruous | I gave Rhonda's dress to Erica. |
| Phrase Structure | Incongruous | I gave Rhonda's to dress Erica. |
| Phrase Structure | Congruous | Jimmy hung Alice's portrait on the wall. |
| Phrase Structure | Incongruous | Jimmy hung Alice's on portrait the wall. |
| Phrase Structure | Congruous | I read Bob's book to my niece. |
| Phrase Structure | Incongruous | I read Bob's to book my niece. |
| Phrase Structure | Congruous | She served Andrew's dish to her grandmother. |
| Phrase Structure | Incongruous | She served Andrew's to dish her grandmother. |
| Phrase Structure | Congruous | Daphne introduced Nick's art to the public. |
| Phrase Structure | Incongruous | Daphne introduced Nick's to art the public. |
| Phrase Structure | Congruous | He turned Susie's bread into sandwiches. |
| Phrase Structure | Incongruous | He turned Susie's into bread sandwiches. |
| Phrase Structure | Congruous | I gave Melissa's syrup to the children. |

| | | |
|---|---|---|
| Phrase Structure | Incongruous | I gave Melissa's to syrup the children. |
| Phrase Structure | Congruous | The prosecutor presented Ron's evidence to the jury. |
| Phrase Structure | Incongruous | The prosecutor presented Ron's to evidence the jury. |

## Appendix G: Sociolinguistic Survey

1. What is your age in years?

2. How do you currently describe your gender identity?

3. What are languages that you speak and level of proficiency?

4. In what state were you born?

5. Approximately how long did you stay in your birth state?

6. Where have you lived previously and for how long?

7. What did you notice about the experiment? Did anything stand out to you while you read and rated the sentences?

Read the following sentences and then answer questions 8a-c:

> *The vegetables need sliced before being cooked.*
> *The silver necklace needs shined before selling.*
> *My laptop needs scanned for virus threats.*
> *The unemployment report needs updated with the latest data.*
> *The kitchen needs cleaned at the end of the day.*

8a) Have you heard anyone say the sentences (listed above) before?

8b) Have you said the sentences (listed above) before?

8c) Who do you expect uses the sentences listed above?

## Appendix H: State Map of Location of Birth for Participants

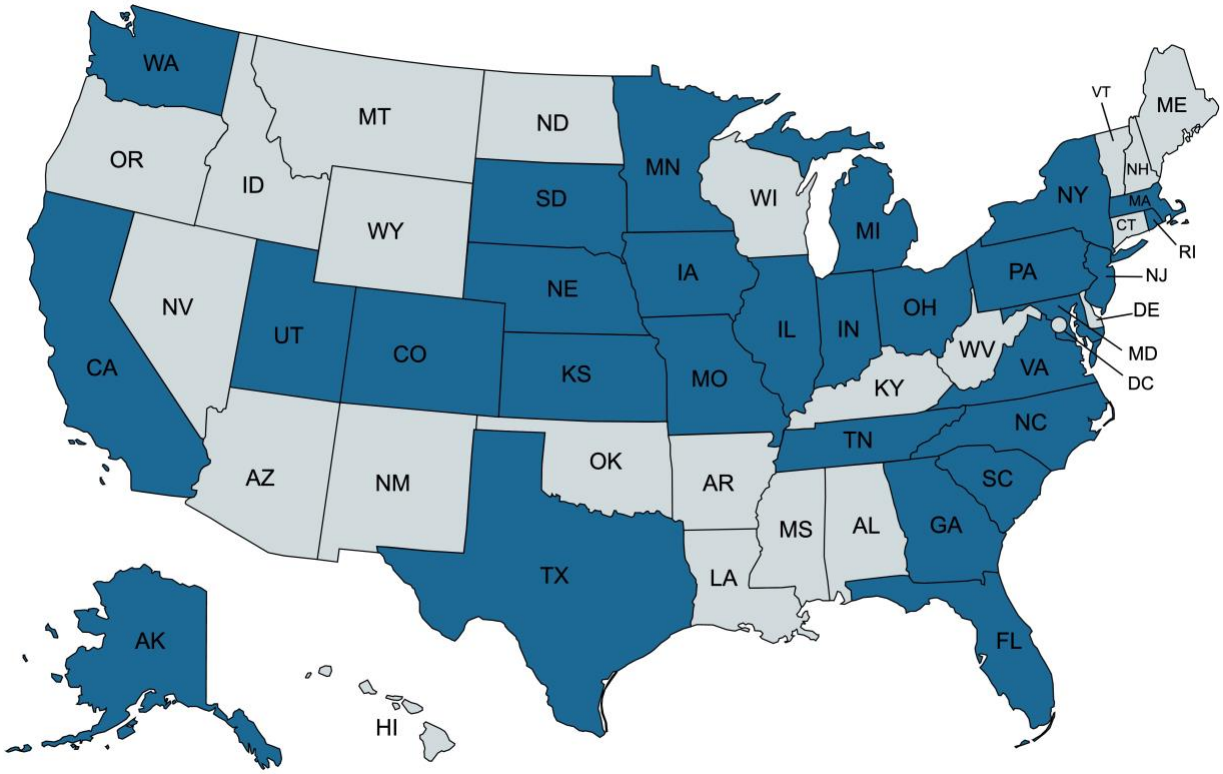


*Figure H-1: Location of Birth for All Participants.* The states colored blue were indicated as birth places of the participants from the 147 participants included in the analysis. The participants came from 28 states with Ohio (n=29) and Pennsylvania (n=40) having the most participants. Participants in the +OHPA group were all born in Ohio or Pennsylvania, except for 4 participants who were born elsewhere (CA, FL, NJ, and NY). Participants in the -OHPA group were from all over the US and the top places were California (n=9), New York (n=7), Texas (n=5). [Image created with mapchart.net].

## Appendix I: Awareness of Self Average Measures in RT and Acceptability

**Average RTs** **Average Acceptability**

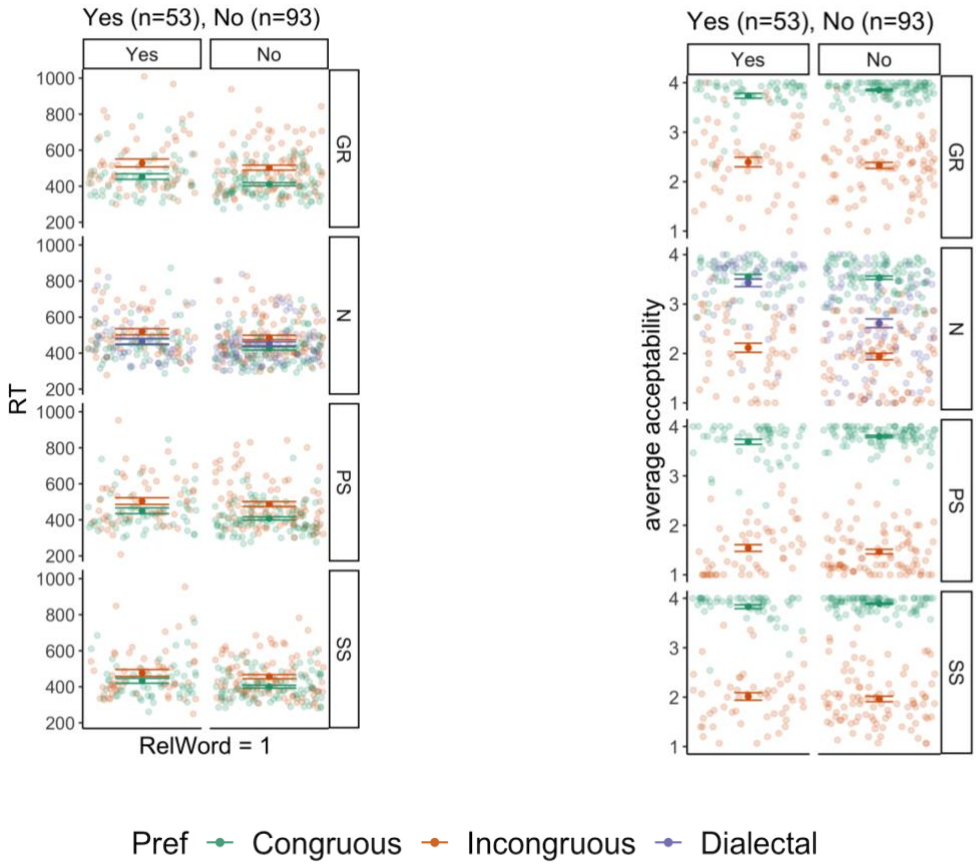


*Figure I-1: Average RTs for Awareness of Self (Right).* In looking at a single point after the target word on the x-axis (Relative word =1), participants answered if they used *needs* sentences with yes (left) or no (right). The preference of each condition are lines are congruous (green), incongruous (orange), and dialectal (purple). RT is listed across the y-axis in milliseconds. The bolded dots represent the average RT of that preference in the *needs* condition for all participants within the group, and the standard error is shown by the error bars for each of the averages. The transparent dots represent the by-subject RT average for the preferences in each condition. *Average Acceptability Judgments for Awareness of Self (Left).* The x-axis is a placeholder to show all preferences in a single condition. Acceptability is listed across the y-axis, where 1 is *unacceptable* and 4 is *acceptable*. The bolded dots represent the average acceptability judgment of that preference in the *needs* condition for all participants within the group, and the standard error is shown by the error bars for each of the averages. The transparent dots represent the by-subject acceptability judgment average for the preferences in each condition. [Yes n=53; No n=93].

## Appendix J: Awareness of Others Average Measures in RT and Acceptability

**Average Reading Time** **Average Acceptability**

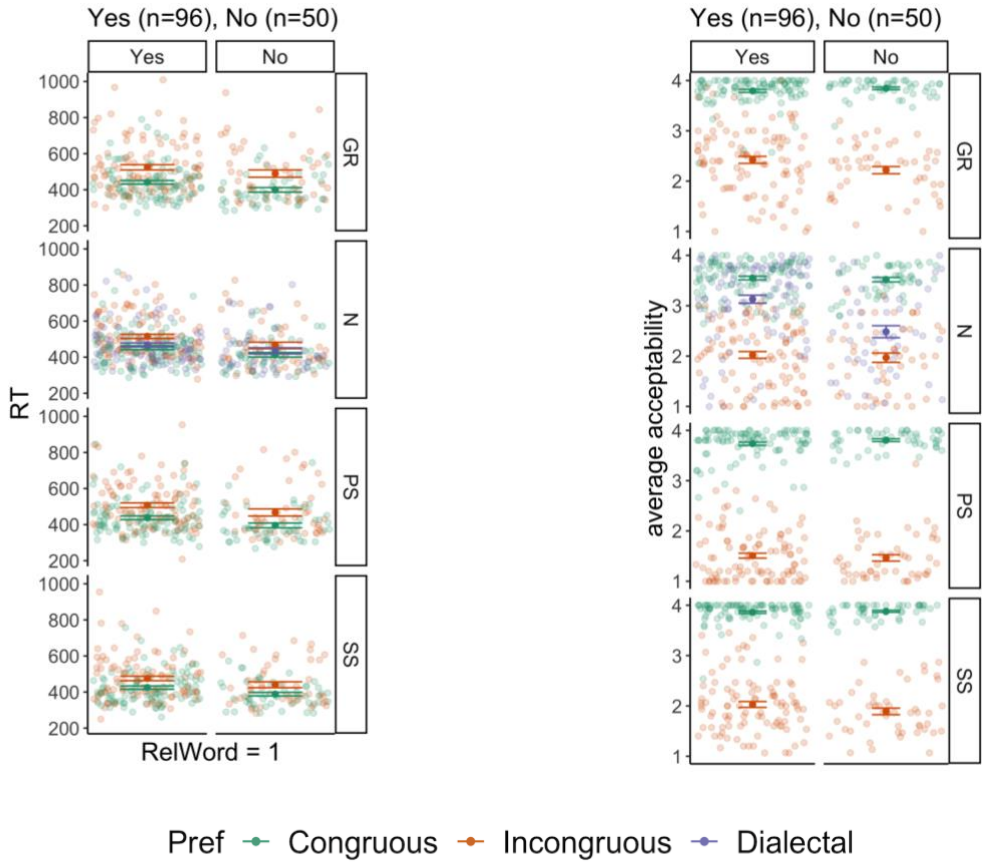


*Figure J-1: Average RTs for Awareness of Others (Right).* In looking at a single point after the target word on the x-axis (Relative word =1), participants answered if others they know used needs sentences with yes (left) or no (right). The preference of each condition are lines are congruous (green), incongruous (orange), and dialectal (purple). RT is listed across the y-axis in milliseconds. The bolded dots represent the average RT of that preference in the needs condition for all participants within the group, and the standard error is shown by the error bars for each of the averages. The transparent dots represent the by-subject RT average for the preferences in each condition. *Average Acceptability Judgments for Awareness of Others (Left).* The x-axis is a placeholder to show all preferences in a single condition. Acceptability is listed across the y-axis, where 1 is unacceptable and 4 is acceptable. The bolded dots represent the average acceptability judgment of that preference in the needs condition for all participants within the group, and the standard error is shown by the error bars for each of the averages. The transparent dots represent the by-subject acceptability judgment average for the preferences in each condition. [Yes n=96; No n=50].

## Appendix K: Top Indexing Themes

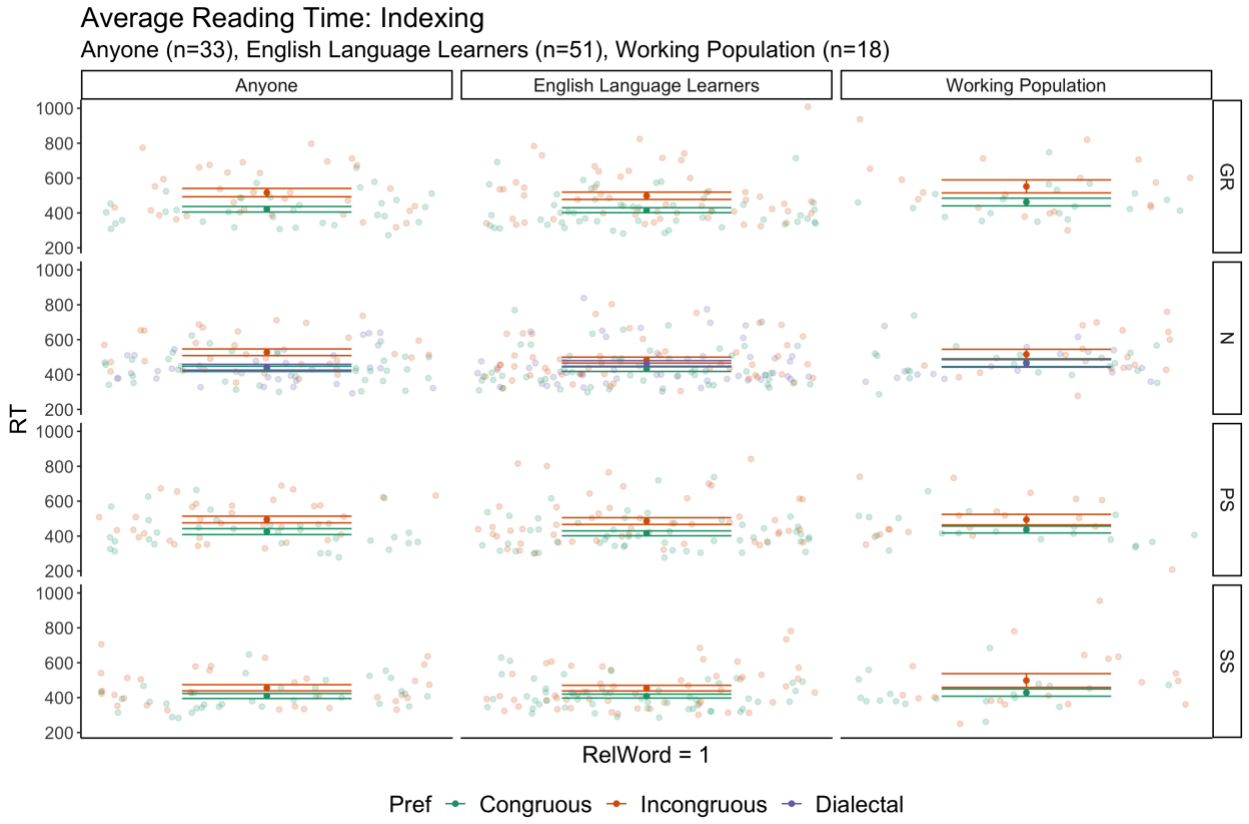


*Figure K-1: Average Reading Time Grouped by Common Indexing Theme in all Conditions.* In looking at a single point after the target word on the x-axis (Relative word =1), participants answered who they think would use *needs* sentences. The top three themes are listed on the top of the graph Anyone (left), English Language Learners (middle), and Working Population (right). The preference of each condition are lines are congruous (green), incongruous (orange), and dialectal (purple). RT is listed across the y-axis in milliseconds. The bolded dots represent the average RT of that preference in the needs condition for all participants within the group, and the standard error is shown by the error bars for each of the averages. The transparent dots represent the by-subject RT average for the preferences in each condition. [Anyone n=33, English Language Learners n=51, Working Population n=18].


*Figure K-2: Average Acceptability Judgment Grouped by Common Indexing Theme in all Conditions.* Participants answered who they think would use needs sentences. The top three themes are listed on the top of the graph Anyone (left), English Language Learners (middle), and Working Population (right). The preference of each condition are lines are congruous (green), incongruous (orange), and dialectal (purple). The x-axis is a placeholder to show all preferences in a single condition. Acceptability is listed across the y-axis, where 1 is unacceptable and 4 is acceptable. The bolded dots represent the average acceptability judgment of that preference in the needs condition for all participants within the group, and the standard error is shown by the error bars for each of the averages. The transparent dots represent the by-subject acceptability judgment average for the preferences in each condition. [Anyone n=33, English Language Learners n=51, Working Population n=18].

## Bibliography

Ainsworth-Darnell, K., Shulman, H. G., & Boland, J. E. (1998). Dissociating Brain Responses to Syntactic and Semantic Anomalies: Evidence from Event-Related Potentials. *Journal of Memory and Language*, *38*(1), 112–130. https://doi.org/10.1006/jmla.1997.2537

Balconi, M., & Pozzoli, U. (2005). Comprehending Semantic and Grammatical Violations in Italian. N400 and P600 Comparison with Visual and Auditory Stimuli. *Journal of Psycholinguistic Research*, *34*(1), 71–98. https://doi.org/10.1007/s10936-005-3633-6

Batterink, L., & Neville, H. J. (2013). The Human Brain Processes Syntax in the Absence of Conscious Awareness. *Journal of Neuroscience*, *33*(19), 8528–8533. https://doi.org/10.1523/JNEUROSCI.0618-13.2013

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

Bresnan, J. (2007). Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In S. Featherston & W. Sternefeld (Eds.), *Roots: Linguistics in search of its evidential base* (pp. 75–96). Mouton de Gruyter.

Bulkes, N. Z., Christianson, K., & Tanner, D. (2020). Semantic constraint, reading control, and the granularity of form-based expectations during semantic processing: Evidence from ERPs. *Neuropsychologia*, *137*, 107294. https://doi.org/10.1016/j.neuropsychologia.2019.107294

Carreiras, M., Garnham, A., Oakhill, J., & Cain, K. (1996). The Use of Stereotypical Gender Information in Constructing a Mental Model: Evidence from English and Spanish. *The Quarterly Journal of Experimental Psychology Section A*, *49*(3), 639–663. https://doi.org/10.1080/713755647

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Praeger.

Chomsky, N. (2015). *Aspects of the theory of syntax: 50th Anniversary Edition* (Fiftieth anniversary edition). The MIT Press.

Coulson, S., King, J. W., & Kutas, M. (1998). Expect the Unexpected: Event-related Brain Response to Morphosyntactic Violations. *Language and Cognitive Processes*, *13*(1), 21–58. https://doi.org/10.1080/016909698386582

Dąbrowska, E. (2010). Naive v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review*, *27*(1), 1–23. https://doi.org/10.1515/tlir.2010.001

Davidson, D. J., & Indefrey, P. (2007). An inverse relation between event-related and time–frequency violation responses in sentence processing. *Brain Research*, *1158*, 81–92. https://doi.org/10.1016/j.brainres.2007.04.082

Delorme, A., Miyakoshi, M., Jung, T.-P., & Makeig, S. (2015). Grand average ERP-image plotting and statistics: A method for comparing variability in event-related single-trial EEG activities across subjects and conditions. *Journal of Neuroscience Methods*, *250*, 3–6. https://doi.org/10.1016/j.jneumeth.2014.10.003

Dröge, A., Fleischer, J., Schlesewsky, M., & Bornkessel-Schlesewsky, I. (2016). Neural mechanisms of sentence comprehension based on predictive processes and decision certainty: Electrophysiological evidence from non-canonical linearizations in a flexible

word order language. *Brain Research*, *1633*, 149–166. https://doi.org/10.1016/j.brainres.2015.12.045

Featherston, S. (2008). The Decathlon Model of Empirical Syntax. In *The Decathlon Model of Empirical Syntax* (pp. 187–208). De Gruyter Mouton. https://doi.org/10.1515/9783110197549.187

Finegan, E. (2019). *What Is "Correct" Language?* What Is "Correct" Language? https://www.linguisticsociety.org/resource/what-correct-language

Francis, E. (2022). *Gradient acceptability and linguistic theory*. Oxford University Press.

Fraundorf, S. H., & Jaeger, T. F. (2016). Readers generalize adaptation to newly-encountered dialectal structures to other unfamiliar structures. *Journal of Memory and Language*, *91*, 28–58. https://doi.org/10.1016/j.jml.2016.05.006

Friederici, A. D., & Meyer, M. (2004). The brain knows the difference: Two types of grammatical violations. *Brain Research*, *1000*(1–2), 72–77. https://doi.org/10.1016/j.brainres.2003.10.057

García, F. M. (2017). *Brain responses to contrastive and noncontrastive morphosyntactic structures in African American English and Mainstream American English: ERP evidence for the neural indices of dialect*. Columbia University.

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

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

Hagoort, P., & Brown, C. M. (2000). *ERP effects of listening to speech compared to reading: The P600/SPS to syntactic violations in spoken sentences and rapid serial visual presentation*. 19.

Hansen, T. (2005). *Auditory and Visual Correlates of the Processing of Gapping Structures in Adults*. Brigham Young University.

Hanulíková, A., Alphen, P. M. van, Goch, M. M. van, & Weber, A. (2012). When One Person's Mistake Is Another's Standard Usage: The Effect of Foreign Accent on Syntactic Processing. *Journal of Cognitive Neuroscience*, *24*(4), 878–887. https://doi.org/10.1162/jocn_a_00103

Hildebrandt, T. (2017). *Accounting for the Variation in the Distribution of Quick and Quickly: A Minimalist Syntactic Analysis* [QRP]. University of Michigan.

Hildebrandt, T., & Brennan, J. R. (2022). *Adaptation effects limited to online measures in the needs past participle construction* [Virtual 5 Minute Talk]. 35th Annual Conference on Human Sentence Processing Conference, University of California Santa Cruz.

Höller, Y. (2021). Quantitative EEG in Cognitive Neuroscience. *Brain Sciences*, *11*(4), 517. https://doi.org/10.3390/brainsci11040517

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*(2), 228–238. https://doi.org/10.1037/0096-3445.111.2.228

Kaschak, M. P. (2006). What this construction needs is generalized. *Memory & Cognition*, *34*(2), 368–379. https://doi.org/10.3758/BF03193414

Kaschak, M. P., & Glenberg, A. M. (2004). This construction needs learned. *Journal of Experimental Psychology: General*, *133*(3), 450–467. https://doi.org/10.1037/0096-3445.133.3.450

Kim, A. E., Oines, L., & Miyake, A. (2018). Individual differences in verbal working memory underlie a tradeoff between semantic and structural processing difficulty during language comprehension: An ERP investigation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(3), 406–420. https://doi.org/10.1037/xlm0000457

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

Kuperberg, G. R., Sitnikova, T., Caplan, D., & Holcomb, P. J. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cognitive Brain Research*, *17*(1), 117–129. https://doi.org/10.1016/S0926-6410(03)00086-7

Kutas, M., & Federmeier, K. D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, *62*(1), 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Kutas, M., & Hillyard, S. A. (1980). Reading Senseless Sentences: Brain Potentials Reflect Semantic Incongruity. *Science*, *207*(4427), 203–205. https://doi.org/10.1126/science.7350657

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

Labov, W. (1996). When Intuitions Fail. In L. McNair, K. Singer, L. Dolbrin, & M. Aucon (Eds.), *Papers from the Parasession on Theory and Data in Linguistics* (Vol. 32, pp. 77–106). Chicago Linguistic Society.

Labov, W., Ash, S., & Boberg, C. (2005). *The Atlas of North American English: Phonetics, Phonology and Sound Change*. Walter de Gruyter GmbH. http://ebookcentral.proquest.com/lib/umichigan/detail.action?docID=453906

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

Leckey, M., & Federmeier, K. D. (2019). The P3b and P600(s): Positive contributions to language comprehension. *Psychophysiology*, e13351. https://doi.org/10.1111/psyp.13351

Leinonen, A., Brattico, P., Järvenpää, M., & Krause, C. M. (2008). Event-related potential (ERP) responses to violations of inflectional and derivational rules of Finnish. *Brain Research*, *1218*, 181–193. https://doi.org/10.1016/j.brainres.2008.04.049

Luck, S. J. (2014). *An introduction to the event-related potential technique* (Second edition). The MIT Press.

Martin, R. (2001). Null Case and the Distribution of PRO. *Linguistic Inquiry*, *32*(1), 141–166. https://doi.org/10.1162/002438901554612

*MATLAB* (9.5.0.1033004). (2019). MathWorks.

McKinnon, R. (1996). Constraints on Movement Phenomena in Sentence Processing: Evidence from Event-related Brain Potentials. *Language and Cognitive Processes*, *11*(5), 495–524. https://doi.org/10.1080/016909696387132

Mehravari, A. S., Tanner, D., Wampler, E. K., Valentine, G. D., & Osterhout, L. (2015). Effects of Grammaticality and Morphological Complexity on the P600 Event-Related Potential Component. *PLOS ONE*, *10*(10), e0140850. https://doi.org/10.1371/journal.pone.0140850

Molinaro, N., Barber, H. A., & Carreiras, M. (2011). Grammatical agreement processing in reading: ERP findings and future directions. *Cortex*, *47*(8), 908–930. https://doi.org/10.1016/j.cortex.2011.02.019

Murray, T. E., & Simon, B. L. (2006). 1. What is dialect? — Revisiting the Midland. In T. E. Murray & B. L. Simon (Eds.), *Varieties of English Around the World* (Vol. G36, pp. 1–30). John Benjamins Publishing Company. https://doi.org/10.1075/veaw.g36.02mur

Neville, H., Nicol, J., Barss, A., Forster, K. I., & Garrett, M. F. (1991). *Syntactically Based Sentence Processing Classes: Evidence from Event-Related Brain Potentials*. *3*(2), 15.

Nevins, A., Dillon, B., Malhotra, S., & Phillips, C. (2007). The role of feature-number and feature-type in processing Hindi verb agreement violations. *Brain Research*, *1164*, 81–94. https://doi.org/10.1016/j.brainres.2007.05.058

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113. https://doi.org/10.1016/0028-3932(71)90067-4

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Computational Intelligence and Neuroscience*, *2011*, 1–9. https://doi.org/10.1155/2011/156869

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., Bersick, M., & Mclaughlin, J. (1997). Brain potentials reflect violations of gender stereotypes. *Memory & Cognition*, *25*(3), 273–285. https://doi.org/10.3758/BF03211283

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

Osterhout, L., Holcomb, P. J., & Swinney, D. A. (1994). Brain Potentials Elicited by Garden-Path Sentences: Evidence of the Application of Verb Information During Parsing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 18. https://doi.org/10.1037//0278-7393.20.4.786

Osterhout, L., McKinnon, R., Bersick, M., & Corey, V. (1996). On the language specificity of the brain response to syntactic anomalies: Is the syntactic positive shift a member of the P300 family? *Journal of Cognitive Neuroscience*, *8*(6), 507–527. https://go.gale.com/ps/i.do?p=AONE&sw=w&issn=0898929X&v=2.1&it=r&id=GALE%7CA387591416&sid=googleScholar&linkaccess=abs

Osterhout, L., & Mobley, L. A. (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

Osterhout, L., & Nicol, J. (1999). On the Distinctiveness, Independence, and Time Course of the Brain Responses to Syntactic and Semantic Anomalies. *Language and Cognitive Processes*, *14*(3), 283–317. https://doi.org/10.1080/016909699386310

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

Polich, J. (2011). *Neuropsychology of P300*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195374148.013.0089

Sassenhagen, J., Schlesewsky, M., & Bornkessel-Schlesewsky, I. (2014). The P600-as-P3 hypothesis revisited: Single-trial analyses reveal that the late EEG positivity following linguistically deviant material is reaction time aligned. *Brain and Language*, *137*, 29–39. https://doi.org/10.1016/j.bandl.2014.07.010

Schütze, C. T. (2016). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology* [Application/pdf]. Language Science Press. http://langsci-press.org/catalog/book/89

Snyder, W. (2000). An Experimental Investigation of Syntactic Satiation Effects. *Linguistic Inquiry*, *31*(3), 575–582. https://doi.org/10.1162/002438900554479

Sorace, A., & Keller, F. (2005). Gradience in linguistic data. *Lingua*, *115*(11), 1497–1524. https://doi.org/10.1016/j.lingua.2004.07.002

Sprouse, J. (2015). Three open questions in experimental syntax. *Linguistics Vanguard*, *1*(1), 89–100. https://doi.org/10.1515/lingvan-2014-1012

Sprouse, J. (2022). Acceptability Judgments. In J. Sprouse (Ed.), *The Oxford Handbook of Experimental Syntax*. Oxford University Press.

Sprouse, J., Schütze, C. T., & Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010. *Lingua*, *134*, 219–248. https://doi.org/10.1016/j.lingua.2013.07.002

Squires, L. (2014a). Social Differences in the Processing of Grammatical Variation. *University of Pennsylvania Working Papers in Linguistics*, *20*(2), 12.

Squires, L. (2014b). Processing, Evaluation, Knowledge: Testing the Perception of English Subject–Verb Agreement Variation. *Journal of English Linguistics*, *42*(2), 144–172. https://doi.org/10.1177/0075424214526057

Squires, L. (2016). Processing Grammatical Differences: Perceiving versus Noticing. In A. M. Babel (Ed.), *Awareness and Control in Sociolinguistic Research* (pp. 80–103). Cambridge University Press. https://doi.org/10.1017/CBO9781139680448.006

Strelluf, C. (2020). Needs+PAST PARTICIPLE in regional Englishes on Twitter. *World Englishes*, *39*(1), 119–134. https://doi.org/10.1111/weng.12451

Swaab, T. Y., Ledoux, K., Camblin, C. C., & Boudewyn, M. A. (2011). *Language-Related ERP Components*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195374148.013.0197

Tanner, D. (2018). *General files for "Robust neurocognitive individual differences in grammatical agreement processing: A latent variable approach"* [Database]. Harvard Dataverse V2. https://doi.org/10.7910/DVN/DKEKBH

Tanner, D. (2019). Robust neurocognitive individual differences in grammatical agreement processing: A latent variable approach. *Cortex*, *111*, 210–237. https://doi.org/10.1016/j.cortex.2018.10.011

Tanner, D., Goldshtein, M., & Weissman, B. (2018). Individual Differences in the Real-Time Neural Dynamics of Language Comprehension. In *Psychology of Learning and Motivation* (Vol. 68, pp. 299–335). Elsevier. https://doi.org/10.1016/bs.plm.2018.08.007

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

Vafaee, P., Suzuki, Y., & Kachisnke, I. (2017). VALIDATING GRAMMATICALITY JUDGMENT TESTS: Evidence from Two New Psycholinguistic Measures. *Studies in Second Language Acquisition*, *39*(1), 59–95. https://doi.org/10.1017/S0272263115000455

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

Van Berkum, J. J. A., van den Brink, D., Tesink, C. M. J. Y., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience*, *20*(4), 580–591. https://doi.org/10.1162/jocn.2008.20054

Vincenzi, M. D., Job, R., Di Matteo, R., Angrilli, A., Penolazzi, B., Ciccarelli, L., & Vespignani, F. (2003). Differences in the perception and time course of syntactic and semantic violations. *Brain and Language*, *85*(2), 280–296. https://doi.org/10.1016/S0093-934X(03)00055-5

Weissler, R. E. (2021). *Leveraging African American English Knowledge: Cognition and Multidialectal Processing* [Thesis]. https://doi.org/10.7302/2706

Weissler, R. E., & Brennan, J. R. (2020). How do Listeners Form Grammatical Expectations to African American Language? *University of Pennsylvania Working Papers in Linguistics*, *25*(2), 9. https://repository.upenn.edu/pwpl/vol25/iss2/16

Zaharchuk, H. A., Shevlin, A., & van Hell, J. G. (2021). Are our brains more prescriptive than our mouths? Experience with dialectal variation in syntax differentially impacts ERPs and behavior. *Brain and Language, 218*, 104949. https://doi.org/10.1016/j.bandl.2021.104949

Zeitlin, M. (2020). *Individual Differences in Linguistic Prediction in Native Language Comprehension and Second Language Learning* [Ph.D., University of Washington]. http://www.proquest.com/docview/2455588862/abstract/BAC07E4DD6C14206PQ/1