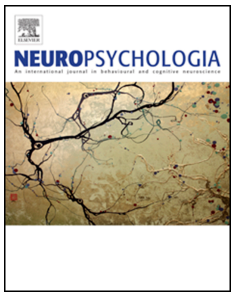

Linguistic Empathy: Behavioral measures, neurophysiological correlates, and correlation with Psychological Empathy

Trevor Kann, Steven Berman, Michael S. Cohen, Emily Goldknopf, Merve Gülser, Gennady Erlikhman, Kristi Trinh, Olga T. Yokoyama, Eran Zaidel

PII: S0028-3932(23)00184-7

DOI: https://doi.org/10.1016/j.neuropsychologia.2023.108650

Reference: NSY 108650

To appear in: *Neuropsychologia*

Received Date: 14 October 2022

Revised Date: 8 May 2023

Accepted Date: 14 July 2023

Please cite this article as: Kann, T., Berman, S., Cohen, M.S., Goldknopf, E., Gülser, M., Erlikhman, G., Trinh, K., Yokoyama, O.T., Zaidel, E., Linguistic Empathy: Behavioral measures, neurophysiological correlates, and correlation with Psychological Empathy, *Neuropsychologia* (2023), doi: https://doi.org/10.1016/j.neuropsychologia.2023.108650.

Linguistic Empathy: Behavioral measures, neurophysiological correlates, and correlation with Psychological Empathy

CRediT Statement:
Trevor Kann (corresponding author, kann@humnet.ucla.edu): Conceptualization, Methodology, Data Collection, Data Curation, Writing – Original Draft, Writing – Review & Editing, Funding Acquisition. **Steven Berman** (sberman@ucla.edu): Software, Analysis, Data Curation, Writing – Original Draft, Writing – Review & Editing. **Michael S. Cohen** (mcohen3@sas.upenn.edu): Software, Analysis, Data Curation, Writing – Original Draft, Writing – Review & Editing. **Emily Goldknopf** (egoldknopf@gmail.com): Analysis, Data Curation, Writing – Original Draft, Writing – Review & Editing. **Merve Gülser** (mervegulser@ucla.edu): Data Collection, Data Curation. **Gennady Erlikhman** (gennady@ucla.edu): Software. **Kristi Trinh** (kltrinh@g.ucla.edu): Data Curation. **Olga T. Yokoyama** (olga@humnet.ucla.edu): Conceptualization, Writing – Original Draft, Writing – Review & Editing, Funding Acquisition. **Eran Zaidel**: Conceptualization, Methodology, Analysis, Resources, Writing – Original Draft, Writing – Review & Editing, Supervision, Project Administration, Funding Acquisition.

Linguistic Empathy: Behavioral measures, neurophysiological correlates, and correlation with Psychological Empathy

Trevor Kann [a *], Steven Berman [b, c], Michael S. Cohen [c], Emily Goldknopf [c], Merve Gülser [c], Gennady Erlikhman [c], Kristi Trinh [c], Olga T. Yokoyama [a], Eran Zaidel [c]

[a] Department of Applied Linguistics, UCLA: 3125 Campbell Hall, Los Angeles, CA 90095, USA

[b] The Semel Institute and Brain Research Institute, UCLA: 760 Westwood Plaza, Los Angeles, CA 90024, USA

[c] Department of Psychology, UCLA: 1285 Psychology Building, Box 951563, Los Angeles, CA 90095, USA

*Corresponding author
Tel.: 562.773.1756
Email: kann@humnet.ucla.edu

Acknowledgments:
The authors are honored to have participated in one of the final empirical projects of Eran Zaidel's career. Data collection and analysis, and early drafts of this manuscript, were completed under Eran's leadership prior to his death. Eran's intense enthusiasm and curiosity, apparent in his approach to scientific inquiry and in his varied interests, inspires us all.

Additionally, we are grateful for the financial support from the UCLA Transdisciplinary Seed Grant.

Declaration of Interests:
☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☒ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Steven Berman, Eran Zaidel, and Trevor Kann report equipment or software was provided by Pacific Development & Technology, LLC.

ABSTRACT

Relations among behavioral, psychological, and electrophysiological correlates of Linguistic Empathy were examined in two experiments using lateralized stimuli. Linguistic Empathy is defined as a linguistic manifestation of the point of view the speaker assumes toward the content of the utterance, and of the speaker's attitude toward/identification with the referents therein. Linguistic choices made by the speaker among multiple logically and referentially synonymous lexical and grammatical options reveal the speaker's perspectives. In experiment 1, acceptability ratings were measured for Context-Target sentence pairs that did or did not violate two Empathy Hierarchies (Person Empathy Hierarchy and Topic Empathy Hierarchy); the Empathy Quotient (EQ) test of Psychological Empathy was also administered. Ratings were lower for sentence pairs that violated both hierarchies than for those violating neither and were intermediate for sentences violating only one hierarchy. Linguistic Empathy (LE) was operationalized as the difference in ratings between sentences violating both vs. neither empathy hierarchy; this measure correlated positively with EQ. Experiment 2 replicated those results with new participants and measured reaction time and EEG during ratings. While there were no effects of hemisphere or visual field on the linguistic variables, the amplitude of a positive event-related potential deflection at 380 ms provided a partial electrophysiological correlate for LE. Its difference measure correlated with behavioral LE but not with EQ. Though preliminary, these experiments show that Linguistic Empathy may share information processing computations with Psychological Empathy and have an electrophysiological correlate.

Keywords: psycholinguistics, linguistic empathy, psychological empathy, event-related potentials

1. GENERAL INTRODUCTION

The term *empathy* refers to distinct theoretical concepts in the fields of linguistics and psychology, and this paper examines the potential overlap between the linguistic notion of empathy and the psychological one. We developed a preliminary measure of Linguistic Empathy (LE), the Linguistic Empathy Acceptability Test (LEAT), and compared it to a measure of Psychological Empathy (PE)[1], the Empathy Quotient (EQ) (Baron-Cohen & Wheelwright, 2004). We also used lateralized stimuli in our experiments to investigate if LE was more associated with one hemisphere. In these experiments, we investigated four related hypotheses: (1) Linguistic Empathy can be operationalized and measured behaviorally, (2) LE depends on PE, such that there is a correlation between ratings on the LEAT and the EQ, (3) there is a neurophysiological ERP correlate of LE (Φ(LE)) that is similar in timing to commonly observed ERP correlates in non-verbal tests of Psychological Empathy (Ibanez et al. 2012; Choi & Watanuki, 2014), and (4) Φ(LE) correlates with PE.

1.1 Psychological Empathy

The psychological concept of empathy aligns with common notions of empathy, and, in this study, refers to the ability to a) understand the thoughts and emotions of others, b) to experience these emotions with others, and c) to respond appropriately in these contexts (Kann, 2017). Psychological Empathy is typically divided into emotional (or affective) empathy and cognitive empathy. Emotional empathy is the prosocial component of Psychological Empathy and is based on the vicarious emotional experience of a person observing, possessing and

[1] In this paper, the concepts of Linguistic Empathy and Psychological Empathy are written in full, and the behavioral and electrophysiological measures of these concepts are written as abbreviations (LE and PE).

responding to common sensory-motor experiences, such as crying, of another person (Bryant, 1982). Thus, experience sharing is perceptual/motor, automatic, and bottom-up. Crucially, *emotional empathy* must involve an emotional response that is socially appropriate (Baron-Cohen & Wheelwright, 2004; Lawrence et al., 2004). In contrast, cognitive empathy does not involve the emotional experience of emotional empathy and instead focuses on the rational and unemotional understanding of the perspectives of others (Kohler, 1929; Mead, 1934). Hogan (1969) describes cognitive empathy as "intellectual or imaginative apprehension of another's condition or state of mind," an approach that is not necessarily prosocial. It is inferential, conscious, and top-down (Mead, 1934; Hogan, 1969; Jolliffe & Farrington 2006). The notion of perspective taking that is involved with both cognitive empathy and Linguistic Empathy, as described below, is foundational to the hypothesis that LE and PE may correlate.

1.2 Linguistic Empathy

In contrast with Psychological Empathy, Linguistic Empathy – a language-based concept – involves lexical and grammatical choices. Linguistic Empathy manifests in the points of view expressed by speakers and understood by listeners of sentences/utterances produced in conversation. Linguistic Empathy has been proposed to determine the speaker's momentary and subconscious choice of expressions of perspective among available grammatically or lexically admissible options, all of which otherwise satisfy logical or referential synonymy conditions. Linguistic Empathy has been proposed as the force behind numerous lexical and grammatical choices in typologically unrelated languages (Kuno, 1987; Yokoyama, 1999, 2019; Oshima, 2007). The definition of Linguistic Empathy in Kuno (1987) ranges from metaphoric (camera positioned on the shoulder of speakers revealing how they see the events and states in a sentence)

to descriptive (speakers taking the perspective of certain participants in the events referred to in the sentence or of the participants in the speech event itself). In Yokoyama's (1986, 2000) cognitive discourse model, Empathy towards a given referent corresponds to the frequency with which the referential knowledge of the given referential expression is contained in the speaker's center of current concern (2000, p. 285). Oshima (2007) suggests that Linguistic Empathy is a universal phenomenon across languages and cultures, while raising the question of whether this universality is a "psychologically primitive notion that reflects a certain psychological construct" (2007, p. 733) or a "theoretical construct that can be derived from (the interaction of) other linguistic factors" (2006, p. 167). The speaker's choice of which participant to empathize with more has been shown to favor participants who place high within certain semantic, syntactic, and pragmatic categories, such as person, animacy, agency, topicality, or salience (Kuno & Kaburaki, 1977; Kuno, 1987; Silverstein, 1981, 2016; Deane, 1992). In terms of empathy, each of these categories is organized into hierarchies: First person has higher claims to Linguistic Empathy than other persons, topics have higher claims than non-topics, and so on.

Because speakers tend to favor sharing empathy perspectives with certain entities in discourse over others, Kuno and Kaburaki (1987) propose Linguistic Empathy Hierarchies (EHs) that reflect an expectation or preference for the appropriate selection of perspective in language. Sentences that violate a Linguistic EH are generally lower in linguistic felicity; i.e., they are perceived by native speakers to sound less natural or well-formed even when they are fully grammatical. For instance, it would be felicitous to utter *I'm marrying a guy I met in England last summer!*, but it would sound less natural for a native speaker to utter, in the same situation, *A guy I met in England last summer is marrying me!* It is important to note that both options are grammatical, refer to the same people, and satisfy the same logical truth conditions (i.e., their

meaning is identical). The odd option violates the Person EH, which states that in choosing subjects for sentences, everything else being equal, it is more natural for the speaker to choose him/herself than another entity. In this case, *a guy I met in England last summer* is a noun phrase in the third person that is used as the subject of the verb *marry* instead of the first person, thus violating the Person EH.

There is another factor that contributes to the oddity of the option *A guy I met in England last summer is marrying me!* The subject noun phrase in this sentence is indefinite, as is seen from the indefinite article *a* (as opposed to the definite article *the*). One central pragmatic difference between the definite and indefinite phrases is that indefinite noun phrases denote elements new to the context/discourse, i.e., those that have not yet been established as the discourse topic or participants in the given discourse universe. This leads us to another factor in determining which sentential elements draw greater empathy: the Topic Empathy Hierarchy. To follow Kuno (1987, p. 210), a speaker's empathy with the discourse Topic is greater than that with a non-discourse Topic. According to this hierarchy, we can expect that a definite noun phrase attracts the speaker's empathy more than an indefinite one, and indeed, *The guy I met in England last summer is marrying me!* is, in the same context, less grating than *A guy I met in England last summer is marrying me!* Uttering these sentences after explicitly establishing a *null context* that shows that the interlocutors are present and ready to engage in verbal interaction (e.g., *Guess what, big news*) raises the coherence level considerably.

The surest way to remove the oddity is by preceding the sentence with a non-null *priming* context that explicitly establishes the groom as the topic in the addressee's mind, as in *Guess what, big news! Remember Basil, the guy I met last summer in England, that handsome aristocrat? Well, Basil is marrying me!* Thus, raising the groom on the scale of Topic EH, now

recapitulating the whole story involving *Basil*, gives him a much better chance to compete with the otherwise powerful effect of the Person EH, which favors first person over other persons.

Thus, in our experiments, we manipulated two of the pragmatic categories associated with Linguistic Empathy as proposed in Kuno and Kaburaki (1977), Silverstein (1981), Kuno (1987), and Deane (1992). The first was the Person EH (also called Speech-Act EH), which states that "The speaker cannot empathize with someone else more than with himself" (Kuno 1987, p. 212). The second was the Topic EH, which states that "Given an event or state that involves A and B such that A is coreferential with the topic of the present discourse and B is not, it is easier for the speaker to empathize with A than with B" (Kuno 1987, p. 210). When the topic (high on the Topic EH) happens to be first person (high on the Person EH), that topic's empathy potential is especially strong. In English[2] sentences with reciprocal verbs (e.g., *meet* or *marry*), it is the participant with the strongest empathy potential that is favored as the grammatical subject. In these experiments we incorporate effects only of these two EHs; other EHs proposed by Kuno & Kaburaki (e.g., Surface Structure EH), and grammatical constructions that encode them (e.g., passivization) would require a different experiment design. Incorporating more Linguistic Empathy-related factors could expand this preliminary version of the LEAT to a more exhaustive measure of Linguistic Empathy in future iterations of the experiment.

1.3 Neurophysiological Correlates of Empathy

[2] This is not true for other languages, especially those with morphological case marking and freer word order, such as Slavic languages (Yokoyama 2000, 2019), Native American languages such as Navajo, Cree, or Jinghpaw (Oshima, 2007), or Japanese (Kuno, 1987; Oshima, 2007). In these languages, while empathy is a major factor affecting linguistic form of the sentence, it does not necessarily affect the choice of the grammatical subject.

Shamay-Tsoory et al. (2009) tested and confirmed the theory that emotional and cognitive empathy are neuroanatomically distinct. Evidence from patients with circumscribed brain lesions, from the pattern of fMRI activation during empathy tasks in normal individuals, and from EEG monitoring during such tasks all confirm the distinction between emotional empathy and cognitive empathy. According to Shamay-Tsoory et al. (2009), emotional empathy involves *experience sharing*, often called *neural resonance* by neuroscientists, and is localized in the frontal sensorimotor cortex (inferior frontal gyrus, SMA), inferior parietal lobule (representing the posterior sensorimotor cortex), medial prefrontal cortex, anterior insula, and amygdala (i.e., "the seat of emotions"). In turn, cognitive empathy, which overlaps with Theory of Mind, includes the medial prefrontal cortex (perspective taking), posterior cingulate cortex, posterior superior temporal sulcus, and temporal parietal junction (Shamay-Tsoory et al., 2009).

1.3.1 ERP Correlates of Empathy

Various ERP components have been associated with aspects of Psychological Empathy. In a review, Ibanez et al. (2012) found that experiments on empathy in which participants are shown pictures of body parts or faces of people in pain characteristically yield two ERP component correlates of empathy: an early automatic frontal negative component (N1 or N100) distinguishing painful from non-painful stimuli, and a later controlled midline maximal centroparietal positive component (P3 or P300) representing pain empathy. The P3 is a well-researched large positive component generated between 300 and 600 ms after low-probability task-related items in a series of stimuli. It includes an earlier frontocentral scalp maximal P3a subcomponent, which is associated with the engagement of attention, and a later centroparietal P3b subcomponent, associated with task-related and memorial processing (Polich, 2007). Ibanez

et al. (2012) found the N1 was modulated by contextual aspects of the pain stimuli and the P3 by task demands. The P3 also distinguished stimuli depicting oneself experiencing pain from stimuli depicting others experiencing pain.

The Late Positive Potential (LPP) is conceptualized as a P3-like positivity, elicited by emotional stimuli and which encodes their motivational salience (Choi and Watanuki, 2014). Choi and Watanuki found that when discriminating between facial expressions, the amplitude of both early (300–600 ms) and late (600–800 ms) LPPs correlated with a measure of empathy as a trait.

The N4 or N400 is a centroparietally distributed negative deflection, peaking approximately 400 ms after a contextually unexpected visual or auditory word (Kutas & Hillyard, 1980; Lau et al., 2013), which has also been connected to empathy. Van den Brink et al. (2012) presented spoken sentences in which the speaker's identity was surprising given the context (e.g., an adult voice uttering "I cannot sleep without my teddy bear"), and showed that (for the first block) N4 effects correlated with self-rated empathy using the EQ. In a passive reading task featuring characters with false beliefs (i.e., a character believing a painting to hang in a hallway when instead it is actually in the kitchen), the correlation between EQ and N4 suggested that individuals with high EQ interpret events according to the character's false beliefs, whereas individuals with lower EQ interpret language according to the truth of the situation (Ferguson et al., 2015).

One of our questions was what role hemispheric specialization plays in Psychological Empathy, and whether that would extend to Linguistic Empathy. While the left hemisphere plays a large role in language processing (Taylor & Taylor, 1990), the right hemisphere is known to be involved in pragmatic aspects of language, emotional processing, and social conventions (Zaidel,

1998; Hecht, 2014). Therefore, we predicted that the right hemisphere should take the lead in Linguistic Empathy.

1.4 Plan of the study

In this study, we introduce a Linguistic Empathy Acceptability Test (LEAT) as a preliminary measure of Linguistic Empathy. Experiment 1 (N = 34) collected ratings on a set of sentence pairs that represented violations of empathy hierarchies and examined their correlations with EQ, a measure of Psychological Empathy. Experiment 2 (N = 20) replicated Experiment 1 in a new sample of participants. It used a better measure of reaction time and included EEG monitoring with a whole head montage to search for ERP correlates of LE. Specifically, we analyzed the relationship between LE and the ERP deflections elicited by the final word of the sentence pair when lateralized to one visual hemifield.

2. EXPERIMENT 1

2.1 Methods

2.1.1 Experimental Design

The experiment investigated the Topic and Person Empathy Hierarchies. As discussed above, the Topic EH states that when a context sentence has introduced an entity, whether it is first, second, or third person, then, in English, the preference or anticipation is for this entity to appear as the subject of the following sentence, henceforth the *target sentence*. The Person EH states that in an unbiased context, a first-person subject has the highest claim to empathy and is the preferred subject of the target sentence. It is helpful to understand the experimental design by

considering the four context/target sentence pairs that are the *defining conditions* of the experiment, described below.

Both of our experiments used the same basic experimental design: the independent variables were Context Sentence (Null, Priming) and Subject of Target Sentence (1st Person, 3rd Person). The primary dependent variable was acceptability rating. We predicted that when an EH is violated, the sentence's acceptability rating will be lower than when the EH is observed. Half the target sentences were preceded by a *null* context sentence, which does not mention either a first- or third-person entity. Because the Topic EH is not engaged with a *null* context, the preference for the subject of the Target Sentence defaults to the Person EH. Thus, when the *null* context was followed by a target sentence with a 1st Person subject, then both the Topic EH and Person EH were observed (T+P+). This combination represents the first defining condition (item List A) of the design of the experiment and it was expected to yield the highest acceptability ratings. The other half of the trials with *null* context sentences were followed by target sentences with Third Person subjects. Those trials violated the Topic EH as well as the Person EH (T-P-). This combination represents the last defining condition (item List D) of the design of the experiment and it was expected to yield the lowest acceptability ratings.

In the other two lists, the *priming* context sentences were expected to further modulate the acceptability of the target sentences. The *priming* context sentences explicitly mention a third-person entity, which elevates the relevance of this entity as the subject of the following target sentence. When a *priming* context was followed by a target sentence with first-person subjects, then the Topic EH was violated but the Person EH was observed (T-P+). This combination represents the second defining condition (item List B) of the design of the experiment and was expected to yield an intermediate rating between A and D. When a *priming*

context sentence was followed by a target sentence with third-person subjects, then the Topic EH was observed but the Person EH was violated (T+P-). This combination represents the third defining condition (item List C) and was also expected to yield an intermediate rating between A and D. The relationship between B and C remains unspecified here. The four defining conditions A, B, C, and D are summarized, with examples, in Table 1.

Stimulus type (Defining conditions)	A	D	B	C
Context	Null	Null	Priming	Priming
Target	1st Person	3rd Person	1st Person	3rd Person
Expected rating	Highest	Lowest	Intermediate	Intermediate
Application to Topic (T) and Person (P) EHs	(T+P+)	(T-P-)	(T-P+)	(T+P-)
Example context sentence	Guess what happened Monday?	Guess what happened Monday?	Let me tell you about Matt.	Let me tell you about Matt.
Example target sentence	I fought Matt.	Matt fought me.	I fought Matt.	Matt fought me.

Table 1. Summary of the stimulus sentences, the experimental conditions they represent, and their expected ratings.

As seen from this table, our hypotheses predicted the highest ratings for A, the lowest ratings for D, and intermediate ratings for B and C. Thus, for A minus D (A-D) to be a valid measure of Linguistic Empathy, the following requirements must be fulfilled. First, A-D should be significantly different from zero, as this measure represents combined effects of both EHs. We make no specific predictions about relative ratings of B with C since they both observe one EH and violate another. We look for a significant Context x Target Subject interaction to confirm our expectation that sentences violating one EH yield ratings intermediate to those for sentences violating neither or both EHs. We additionally expect that ratings for B and C stimuli will not be more extreme than ratings for A and D stimuli ("range criterion"); because null effects are acceptable for this criterion, we only planned to run follow-up t-tests when values for B and C stimuli were more extreme numerically than those for A or D, to confirm whether any such differences reach statistical significance. Finally, we hypothesized that individuals with greater EQ will have more sensitivity to linguistic violations of EHs, so individual differences in LE should correlate with individual differences in EQ.

2.1.2 Participants

Thirty-four UCLA undergraduate students (19 females, 15 males) participated in the experiment in exchange for course credit. One additional participant was tested but did not complete the experiment due to equipment failure. All participants were classified as right-handed based on self-report and observation of right-hand dominance for writing. All participants ranged from 18 to 22 years of age and completed an intake survey that assessed linguistic and cultural background (see Appendix). Only native speakers of English were

accepted. Autism Spectrum Conditions and IQ were not screened for. All participants were concurrently enrolled undergraduates at UCLA.

2.1.3 Stimuli

The experimental session contained a mixture of trials intended to manipulate a number of different variables related to Linguistic Empathy, some of which are not relevant to this paper[3]. Data from two "sub-experiments" are most relevant to our questions of interest: One subset consisted of three item groups (48 trials) using reciprocal verbs, and the other consisted of three item groups (48 trials) in which Active/Passive constructions were contrasted. Although the Active/Passive constructions manipulated EHs identically to the Reciprocal sentences, the Active/Passive sentences were not included in the analysis for three primary reasons: 1. The syntax and the word count differ from one construction to the next (e.g., *I like Walter, Walter is liked by me*), 2. Although the passive voice is grammatical, there is well-documented bias against its use in formal settings, and 3. The preliminary analysis disclosed that the Reciprocal sentences yielded a stronger estimate of LE than did the Active/Passive sentences. Consequently, for measuring LE, we only examine the results from the Reciprocal sentences.

The Reciprocal sentences included item groups with the following three verbs: *fought, (finally) met,* and *dated.* Eight trials in each item group had a first-person pronoun as the

[3] The set of 208 items presented to each participant was composed of 12 different item groups, where an item group includes all possible sentence variants presented with a particular verb. These 12 item groups consisted of three types of items: reciprocal verbs, active/passive verb structures and sentences with possessive phrases. The items with possessives manipulated different variables that do not map onto the variables of the other two groups, and they are not discussed here. In addition, the Reciprocal and Active/Passive sets each included one item (on an exploratory basis) that manipulated separate variables from the other items in that group. Those items were also excluded from the analysis.

grammatical subject and a single-syllable proper name as the object, e.g., *I fought Matt*. The subject-object choice was reversed for the other eight trials, with a third person grammatical subject and a first-person object (e.g., *Matt fought me*).

Half of the trials were preceded by a *null* context sentence, and the other half were preceded by a *priming* context sentence, as described above. Context length (within five characters) was controlled across sentences within each set. Trial order was randomized for each participant to avoid possible biases from neighborhood effects from surrounding trials.

2.1.4 Procedure

Participants first completed the aforementioned inventory for assessing linguistic and cultural background (see Appendix), followed by the EQ test (Baron-Cohen & Wheelwright 2004) as a measure of Psychological Empathy. During the LEAT, participants were positioned with their chins on a chinrest and their eyes 57.3 cm from the screen. At that distance, one degree of visual angle is equal to one centimeter of distance on the screen. Participants responded to each trial by pressing one of four keys to rate the level of acceptability of the target sentence. The four acceptability levels were indicated by four illustrations of faces (big smile, slight smile, slight frown, and large frown) and they were shown on the last screen at the end of each trial to remind the participants of the assignment of keys to rating choices. Illustrations were used so that participants were not influenced by subjective words (e.g., “good, great, bad, terrible”) or jargon (e.g., “grammatical/ungrammatical,” “felicitous/infelicitous”), and so minimal language processing was required between trials. The choice of faces as an illustrative representation is consistent with guidance by Toepoel et al. (2019). This display was visible on the screen until one of the response buttons was pressed (Figure 1). Participants responded with the hand

ipsilateral to the visual field of the target; thus, if the target was flashed to left visual hemifield, they had to respond with the left hand using the keys "x", "d", "f", or "v", corresponding to "worst", "bad", "good", and "best" ratings, respectively. Similarly, if the target was flashed to right visual hemifield, the participant had to respond with the right hand using the keys, "b", h", "j", or "m", again corresponding to "worst", "bad", "good", and "best" ratings, respectively. In this schema, participants used the pinky finger for "best" with the left hand and the index finger for "best" with the right hand, and so on. Participants were instructed to indicate how "natural or acceptable" they judged the target sentence to be. Here, the left-most "worst" represented the lowest rating (1), and "best" represented the highest rating (4).

We assumed that, overall, the index finger would produce the fastest responses among the four responding fingers, and that the pinky would produce the slowest responses among the four. The data in the Results section was the average of the two hemispheres (left hemisphere = right visual hemi-field — right response hand; right hemisphere = left visual hemi-field — left response hand). By making the ratings of "best" to "worst" sequential, each hand used different fingers for the ratings, which neutralized/counterbalanced the two opposing effects due to the opposite effect of response fingers in the left and right hand.

The experimental session began with a short practice block of 20 trials using 3 verbs ("hit", "liked", "hated") that manipulated the target EHs using active and passive sentences in order to acclimate participants to the presentation and rating process. Participants were instructed to rate sentences on how "natural and acceptable" the sentences would sound coming from a native speaker of English (i.e., linguistic felicity and pragmatic well-formedness). Specifically, they were instructed "*not* [to] rate the sentences based on their perceived "grammaticality" since all sentences are grammatically correct." The practice block was followed by two experiment

blocks, each containing 104 trials. Of the total of 208 experimental trials, 48 trials were the reciprocal sentences that are analyzed below. Figure 1 shows the structure of each trial. The trial began with a fixation cross flashed in the center of the screen for 50 ms, and participants were instructed to fixate the cross throughout the trial. Next, a context sentence appeared centrally for 2000 ms above the fixation cross, and participants were required to read it silently. The sentence then disappeared, while a brief (50 ms) fixation cross remained. Next, the first part of the target sentence appeared centrally for 1500 ms, but with a blank line standing for the critical final word. Finally, the target word that completed the sentence was flashed in one visual hemifield for 180 ms, at an eccentricity of one degree of visual angle measured at the edge closer to fixation. The word subtended 2-5 degrees. Immediately after the target, an image appeared on the screen, reminding the participants of the response arrangements. This screen remained until the response was given. Lateralization of the target word was intended to probe hemispheric specialization for the different types of stimuli.

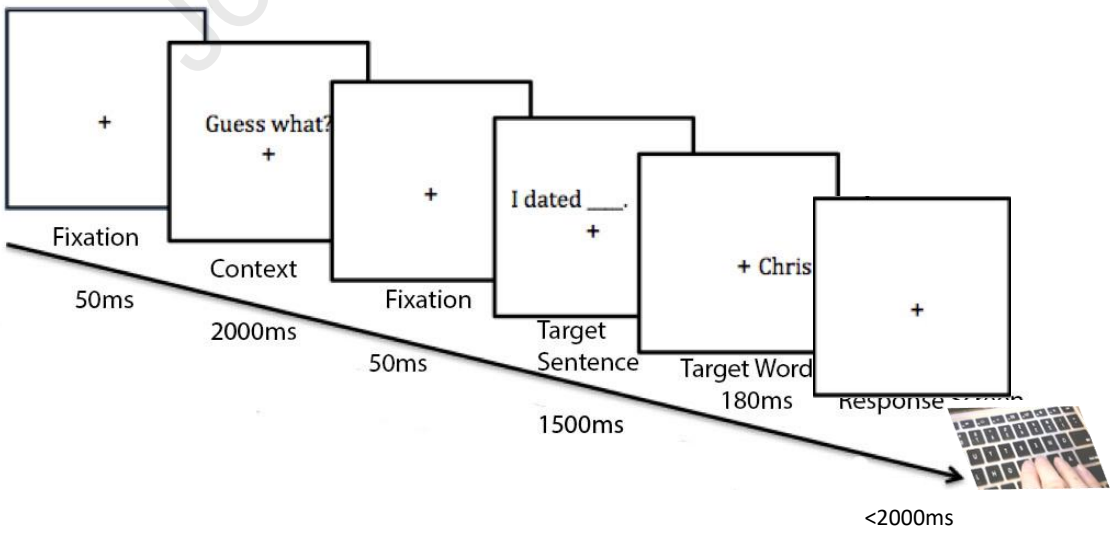

Figure 1. This figure illustrates the screens presented to participants during the experiment. This sample trial consists of a type A stimulus. Since the target is flashed on the right side of the screen, the response to this item should be with the right hand (left hemisphere).

In each trial, the presentation of the lateralized target was followed by a screen that reminded the participant of possible responses; the reminder consisted of a drawing of the positions of the four choices relative to the responding index finger and remained on the screen until the participant responded. Because the time to observe and read the reminder screen was combined with the time to respond, we did not use reaction time as a dependent variable in this first experiment.

2.2 Results

2.2.1 Absence of Laterality Effects

We initially ran a 2 x 2 x 2 ANOVA, Hemisphere x Context (Null, Priming) x Grammatical Subject (1st Person, 3rd Person). This analysis showed a trend towards a main effect of hemisphere, $F(1,33) = 3.75$, $p = .061$, $\eta_p^2 = .10$, with somewhat higher ratings for stimuli presented in the right visual field. However, there were no significant first or second order interactions involving hemisphere, all Fs < 1.19, all ps > .28. Thus, we collapsed across hemisphere in all subsequent analyses.

2.2.2 Control of Effects of the Experimental Sequence

Although the experiment controlled for the frequency of the previous target category (A, B, C, or D), we also examined whether the rating of the previous trial had an effect on the rating of the current one. No significant effect was observed either in this experiment or in Experiment 2. That means that the ratings of each target category in a given trial were not affected by the ratings of the previous trial in the test.

2.2.3 EQ Scores

EQs were obtained for 32 of 34 participants (19 females, 13 males) and ranged from 25-73 out of a possible range of 0-80, with a mean total score of 46.8. Two participants' EQs were not included after they admitted to intentionally distorting their self-report results. The mean score for males was 42.7 and for females was 49.7; this gender difference was not significant, $t(30) = 1.81$, $p = .08$, although the apparent trend is in a direction consistent with prior literature (Baron-Cohen & Wheelwright, 2004).

2.2.4 Analysis of Acceptability Ratings

We first observed that the ratings difference measure, $(A-D)_{Rating}$, was significantly above zero, $M = 1.05$, $t(33) = 8.89$, $p < .001$, $d = 1.52$. Additionally, B and C were both between the values of A and D, satisfying the range criterion. We then ran a 2 x 2, Context (Null, Priming) x Grammatical Subject (1st Person, 3rd Person), ANOVA, with mean acceptability rating as the dependent variable. The critical Context x Grammatical Subject interaction was significant, $F(1, 33) = 16.11$, $p < .001$, $\eta_p^2 = .33$ (Figure 2). We also found a main effect of Grammatical Subject, such that target sentences with 1st Person subjects were rated higher than sentences with 3rd Person subjects ($F(1, 33) = 62.21$, $p < .001$, $\eta_p^2 = .65$), and a main effect of context, with sentences

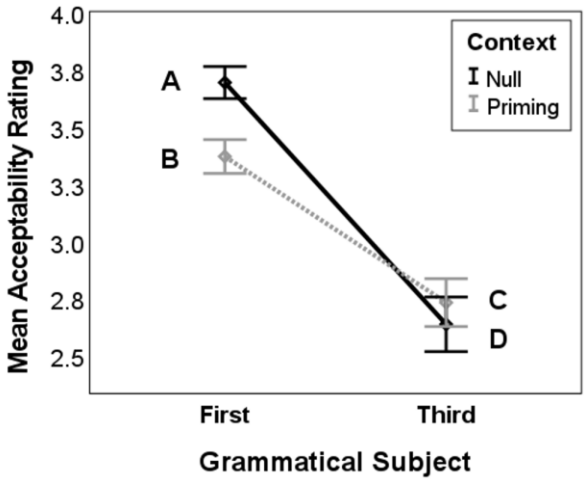

Figure 2. Mean acceptability rating for reciprocal sentences with first person versus third person grammatical subjects, with null or priming context sentences.
Error bars represent +/- 1 SE.

in Null contexts rated higher than those in Priming contexts, $F(1, 33) = 7.72$, $p = .009$, $\eta_p^2 = .19$. Thus, the prerequisite conditions for initially validating a candidate measure of LE were satisfied.

As shown in Figure 3 and Table 2, the mean rating of items in List D correlated negatively with EQ, such that higher EQs were associated with lower ratings, whereas the mean rating of items in List A did not correlate with EQ. The ratings difference measure $(A-D)_{Rating} = (\overline{A}_{Rating} - \overline{D}_{Rating})$ (≡ LE) showed a highly positive correlation with EQ across participants, such that participants with higher EQ had a greater difference between their average ratings for A and D.

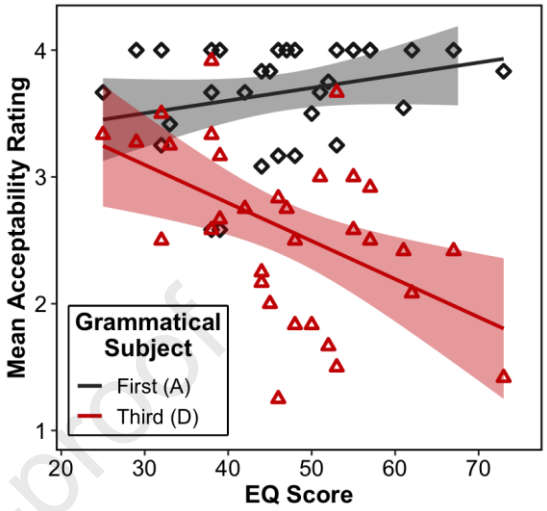

Figure 3. Correlation between Empathy Quotient (EQ) and mean acceptability rating for reciprocal, null context sentences with first person versus third person grammatical subjects. Shaded regions represent 95% confidence intervals.

List	Mean Rating	$r_{between}$ (Rating, EQ)	P	
D	2.59	-.496	.004	**
A	3.67	.268	.137	ns
(A−D)	1.08	.647	< .001	***

Table 2. N = 32: Pearson linear correlation coefficients between ratings and EQ across participants for lists A and D, as well as for the ratings difference measure (A-D) (= LE). *: .05 > p> .01, **: .01 > p > .001, ***: p < .001

2.3 Discussion

Experiment 1 showed that the LEAT satisfied our experimental prerequisites (reliable A-D difference, B and C intermediate to A and D, and a Context x Target Subject interaction) to suggest that A-D was a valid measure of LE. In other words, ratings were lower for sentences that *violated* the EHs than for sentences that *observed* them. Additionally, individuals with higher EQ also had higher LE, showing a greater difference in ratings between A and D items. Average ratings across participants for D items correlated negatively with EQ, such that people with higher EQ rated D items lower. This supports our hypothesis that individuals with higher EQ would have greater sensitivity to violations of Linguistic Empathy. Because all target sentences were grammatically correct, individuals with lower EQ may have focused more on the semantic and syntactic validity of the sentences, and less on the adherence to or violation of EHs.

3. EXPERIMENT 2 (N = 20)

3.1 BEHAVIOR

3.1.1 Introduction

All participants of Experiment 2 (N = 30) had their ongoing EEG recorded while they performed the LEAT. Some participants, however, had unanalyzable EEG data or did not meet the participant criteria (two due to equipment failure, six due to excessive artifact leading to ten or fewer valid trials being present in at least one of the four primary conditions, and two due to exceeding the target age range and not being undergraduates at UCLA).[4] The remaining

[4] The two participants excluded for age were 33 and 34 while the rest of the participants were 18-22 years old. They were also the only 2 participants who were not undergraduate students at UCLA.

participants (N = 20) were retained for ERP analyses. We will report first the behavioral data for these participants followed by the ERP data.

3.1.2 Methods

3.1.2.1 Participants

All participants were classified as right-handed based on self-report and observation of right-hand dominance for writing. Similar to Experiment 1, as determined by the linguistic and cultural inventory (see Appendix), all participants were native speakers of English from 18 to 22. The 20 participants included 14 female and 6 male UCLA undergraduates.

3.1.2.2 Stimuli

The critical stimuli were structured similarly to those used in Experiment 1, although the specific items were different. There were a total of 576 stimuli, which included 6 reciprocal, transitive verbs (*cuddled, dated, fought, high-fived, married, met*). The stimuli also included 12 active/passive verbs, but due to the same lexical and syntactic reasons given for Experiment 1, only the reciprocal verbs are analyzed here, with the exception of one analysis described further below. Each verb was repeated 32 times throughout the stimulus set. Of these items, half were either reflexive (*I met myself*) or did not have a first-person subject or object (*John fought Jake*); these items were not analyzed but were included so that the final word could not be predicted by the sentence preceding it. Additionally, including the reflexive condition introduced sentences that were semantically infelicitous (e.g., *I met myself*) to contrast with the trials with EH violations, ensuring that participants could not predict felicity of a target sentence prior to the final word. Thus, there were 16 repetitions that were analyzed for each verb, yielding 96 trials in

total (24 trials per condition) across the 6 reciprocal verbs. The analyzed independent variables were grammatical subject (1st person vs. 3rd person) and context (null vs. priming); each item appeared four times, twice in each visual field.

3.1.2.3 Procedure

The same parameters for stimuli and timing used in Experiment 1 were also implemented in Experiment 2. However, the response assignment of keyboard positions to rating choices were reversed. Thus, the same fingers were assigned to the same keyboard positions, but for the right hand, "b" was now the highest rating (= 1) and "m" was the lowest rating (= 4). For data presentation in this paper, the numerical rating choices have been reversed to be consistent with Experiment 1. Stimuli were presented using Psychtoolbox (Matlab) instead of E-Prime, and EEG recordings were taken during the LEAT.

3.1.2.4 Analyses

Outlier trials with reaction times that were < 50ms (interpreted as impulsive, anticipatory responses), or trials in which no response was given within the 8 second response window, were excluded from analyses of behavioral data (3.7% of trials in total) in Experiment 2.

In addition to running separate ANOVAs on ratings and on RTs, we correlated EQ with ratings and with RTs across participants for conditions A, D, and A-D. We also correlated ratings with RTs both across and within participants. Correlations across participants indicate average overall relationships between ratings and RTs and emphasize individual differences. By contrast, correlations within participants express the variability of the relationship within individuals

between ratings and RTs for certain items and examine that variability for the different stimulus conditions A and D.

3.1.3 Results

3.1.3.1 Absence of Laterality Effects

A 2 x 2 x 2 (Hemisphere x Context X Grammatical Subject) ANOVA showed no first-order or second-order interaction effects with hemisphere on ratings, all *F*s < 1. There was also no reliable main effect of hemisphere, $F(1,19) = 1.75$, $p = .20$, $\eta_p^2 = .08$, and the weak trend that was present was in the opposite direction as that obtained in Experiment 1. An analogous analysis for reaction times similarly showed no reliable main effects or interactions involving hemisphere, all $F < 2.03$, all $p > .17$. Thus, we again collapsed across hemisphere in subsequent analyses.

3.1.3.2 EQ Scores

EQs were obtained for the 14 females and 6 males. Scores ranged from 23–67 out of a possible range of 0–80, with a mean overall score of 44.95. The mean score for males was 43.17, and for females was 45.71; this gender difference was not significant, $t(18) < 1$.

3.1.3.3 Ratings

3.1.3.3.1 Analyses of Acceptability Ratings

As in Experiment 1, we found a significant $(A-D)_{Rating}$ difference, $M = .38$, $t(19) = 3.75$, $p = .001$, $d = .84$. The range criterion was fulfilled, since the mean ratings for B and C items were between those for A and D. Finally, a 2 x 2 ANOVA showed that the interaction between

Context (Null, Priming) and Grammatical Subject (1st Person, 3rd Person) was significant, $F(1, 19) = 6.521$, $p = .019$, $\eta_p^2 = .26$; see Figure 4. There was a main effect of Grammatical Subject, $F(1, 19) = 7.663$, $p = .012$, $\eta_p^2 = .29$, with higher ratings for target sentences with 1st Person as compared to 3rd Person subjects, but no main effect of Context, $F(1, 19) = 1.098$, $p = .308$, $\eta_p^2 = .05$.

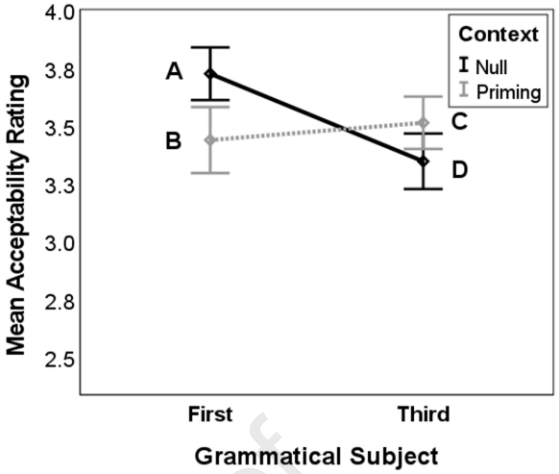

Figure 4. Mean acceptability ratings for target sentences with First Person and Third Person subjects in Null and Priming contexts. Error bars represent +/- 1 SE.

3.1.3.3.2 Ratings and correlations with EQ across participants

The across-participants ratings data are summarized as part of Table 3. EQ was positively correlated with the difference measure $(A\text{-}D)_{Ratings}$ such that higher EQs were associated with a greater difference in ratings and hence greater LE, and negatively correlated with ratings for List D, such that higher EQs were associated with lower ratings for that list (see Figure 5 and Table 3). EQ did not correlate with ratings for List A targets.

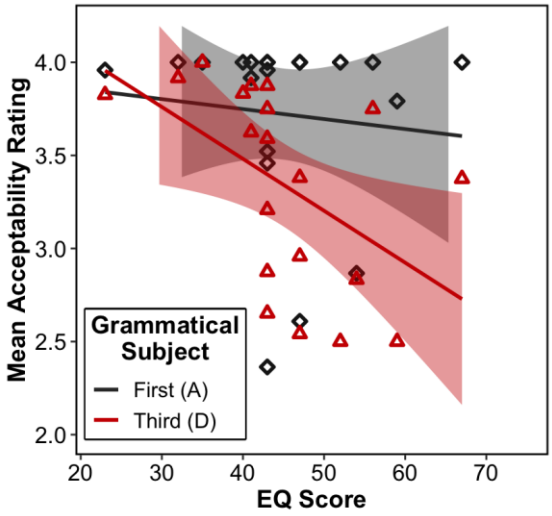

Figure 5. Correlation between Empathy Quotient (EQ) and mean acceptability rating for reciprocal, null context sentences with first person versus third person grammatical subjects. Shaded regions represent 95% confidence intervals.

3.1.3.4 Reaction Times

3.1.3.4.1 Analyses of Reaction Times

As we observed with acceptability ratings, $(A-D)_{RT}$ showed a significant difference, M = -.32, $t(19)$ = -4.70, p < .001, d = -1.05. Additionally, RTs satisfied the range criterion, as mean RTs for B and C items were between A and D mean RTs. Finally, a 2 x 2 ANOVA found that the interaction of Context (Null, Priming) x Grammatical Subject (1st Person, 3rd Person) was significant, $F(1, 19) = 9.024$, p = .007, η_p^2 = .32; see Figure 6. There was also a main effect of Context, with shorter RTs for the Null context, $F(1, 19) = 8.568$, p = .009, η_p^2 = .31, and a main effect of Grammatical Subject, $F(1, 19) = 10.336$, p = .005, η_p^2 = .35, with shorter RTs for target sentences with 1st Person as compared to 3rd Person subjects. Thus, RTs satisfied the same criteria as acceptability ratings for measuring Linguistic Empathy.

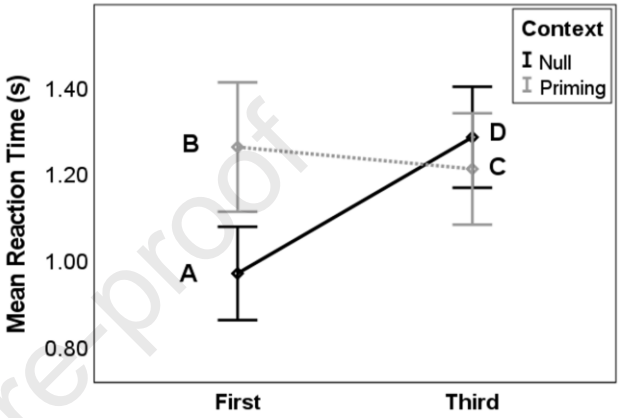

Figure 6. Mean reaction time for target sentences with First Person and Third Person subjects in Null and Priming contexts. Error bars represent +/- 1 SE.

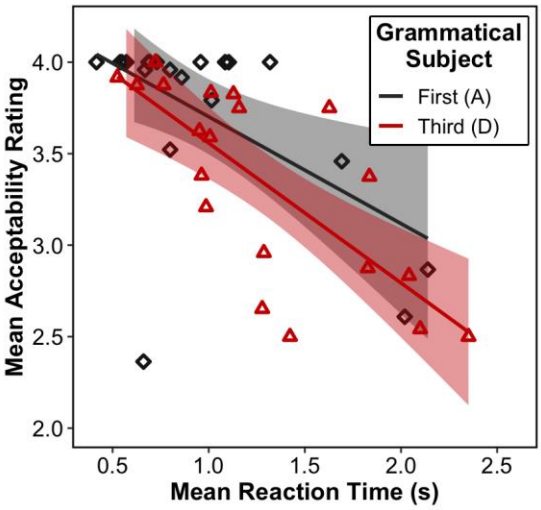

Figure 7. Correlation between mean acceptability ratings and reaction times for reciprocal, null context sentences with first person versus third person grammatical subjects. Shaded regions represent 95% confidence intervals.

Note that the reversed sign for RTs relative to rating measures reflects that D items produce slower RTs than A items, as would be expected. This also causes (see Table 3) the correlation of $(A - D)_{RT}$ with EQ to have the opposite sign as the correlation for

ratings, though the effect in this case was only a trend. Finally, we find that the mean ratings and mean RTs for each measure (A, D, and A – D) were negatively correlated with each other across participants (Figure 7), supporting our contention that these measures are different ways of assessing the same construct.

List	$\overline{Rating}$	$\overline{RT}$	r between($\overline{Rating}$, EQ)		r between($\overline{RT}$, EQ)		r$_{between}$(Rating, RT)	
			r	p-value	r	p-value	r	p-value
D	3.34	1.28 s	-.503	.024 *	.611	.004 **	-.746	< .001 ***
A	3.72	0.97 s	-.102	.669	.394	.086 ~	-.555	.001 **
$(A - D)$	.38	-.32 s	.482	.031 *	-.429	.059 ~	-.569	.009 **

Table 3. Across-participants correlations of mean rating ($\overline{Rating}$) with EQ, of mean reaction time ($\overline{RT}$) with EQ, and of ($\overline{Rating}$) with ($\overline{RT}$). In the last row, $\overline{Rating}$ and $\overline{RT}$ indicate the difference measures (A–D)$_{Rating}$ = $\bar{A}_{Rating} - \bar{D}_{Rating}$ and (A–D)$_{RT}$ = $\bar{A}_{RT} - \bar{D}_{RT}$.
Significance levels are indicated as follows: ~: .10 > p > .05, *: .05 > p> .01, **: .01 > p > .001, ***: p < .001

3.1.3.4.2 Within-participant Correlations of Ratings with their Reaction Times (RatingRT) and their Correlation with EQ

Finally, we examined within-participant correlations of acceptability ratings (Rating) with reaction time (RT) (which we refer to as RatingRT). This provides information about how each participant processed the stimuli. Because LE is defined as $(\overline{A - D})_{Ratings}$ = ($\bar{A}_{Ratings} - \bar{D}_{Ratings}$), we first analyzed within-participants correlations of Rating with RT for A items and for D items. The within-participants correlation of Ratings and their RTs for a given list, RatingRT, was only computed for lists on which a participant's responses showed sufficient variability to determine a valid correlation. Because A items receive high acceptability ratings, creating a ceiling effect, only 6 subjects generated RatingRT for A items, whereas 17 subjects generated RatingRT for D items.

D-item RatingRT correlations were subsequently *z*-transformed to enable further statistical analyses.[5] A one-sample t-test indicated that within-individual rating-RT correlations were negative on average across the sample, $t(16) = -4.06$, $p < .001$, $d = -.98$. Figure 8 illustrates the correlation between EQ and z-transformed within-subject RatingRT correlations for D items, i.e., $r_{between}$(EQ , $RatingRT_{D\ items}$). The correlation was significant ($r = .56$, $p = .021$), indicating that lower EQ was associated with a stronger negative correlation (lower r) between the acceptability ratings and the speed with which they were generated.

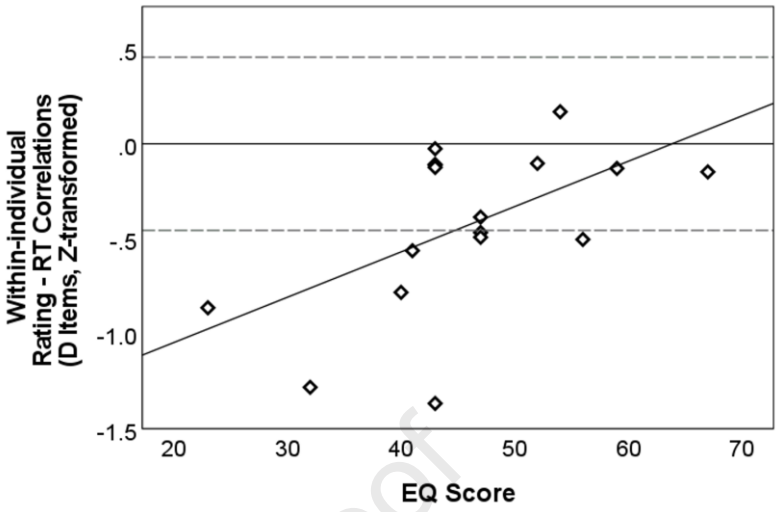

Figure 8. N = 17: Z-transformed across-participants correlation of EQ with within-subject RatingRT for D items. RatingRT correlations outside the dashed lines were significant within individuals (Critical values for $p < .05$: $r = \pm .46$).

3.1.4 Discussion

Similar to Experiment 1, Experiment 2 also found that $(A\text{-}D)_{Rating}$ was a valid measure of LE, and that across participants it correlated positively with EQ. In Experiment 2, we also demonstrated that $(A\text{-}D)_{RT}$ met the criteria to be a measure of LE, and that it was reliably associated with $(A\text{-}D)_{Rating}$. We found marginal evidence suggesting that $(A\text{-}D)_{RT}$ correlated with

[5] Because the bounded nature of raw correlation coefficients violates the assumptions of a parametric test, Fisher's z-transformation was used to prepare the coefficients for the analysis.

EQ as well. Finally, within-participants, the degree to which ratings for D items correlated negatively with reaction times was itself correlated positively with EQ across participants. This "correlation of correlations" suggests that even when individuals with low EQ successfully identify D items as less acceptable, they may have difficulty doing so, leading to greater slowing of RTs.

3.2 ELECTROPHYSIOLOGY

3.2.1 Introduction

In the behavioral findings of Experiments 1 and 2, we found that $(A\text{-}D)_{Rating}$ met our criteria for a valid measure of Linguistic Empathy because (1) it showed an interaction between Context and Target Subject, (2) it satisfied the range criterion, and (3) A-D was significantly different from zero. Our second hypothesis was also borne out: $(A\text{-}D)_{Rating}$ = LE was correlated with EQ. In Experiment 2, we found that RTs fulfilled many of the same criteria, and we explored the correlations among Ratings, RTs and EQ. Using data from Experiment 2, we will now examine whether there is an electrophysiological isomorph for LE that fits the validity criteria and correlates with behavioral measures of LE (ratings and their RTs) and EQ. As discussed in the general introduction, aspects of Psychological Empathy have been associated with a number of ERP components. Complex linguistic processes would be expected to be associated with later cognitive ERPs like the P3b and N400. Because both occur at a similar time (300-600 ms) and scalp distribution (centroparietal midline), and both have been previously associated with Psychological Empathy (Ibanez et al. 2012; Choi and Watanuki, 2014; Van den Brink et al. 2012; Ferguson et al., 2015), we hypothesized that violations of empathy hierarchies

would either increase scalp recorded negativity (N400) or positivity (P3b) at those times and electrodes.

3.2.1.1 Procedures

Six external electrodes were applied to the outer canthus of both eyes, above and beneath the right orbit, and on each mastoid bone. Electroencephalographic activity was digitized and recorded at 512 Hz using BioSemi 64-channel Ag/AgCl electrode caps following International 10-20 electrode placements, and amplified using BioSemi opto-isolated amplifiers. Digital trigger pulses were embedded into the EEG at the onset of each final word. Stimulus codes indicated the visual field of presentation (Left, Right), Grammatical Subject (1st Person, 3rd Person), Context (Null, Priming) and Structure (Active/Passive, Reciprocal).

EEG files were imported into BrainVision Analyzer 2 (www.brainproducts.com) for transformation and analysis. Vertical electrooculogram (VEOG) was measured from above to below the right orbit. Horizontal electrooculogram (HEOG) was measured from the left to right outer canthus. The other channels were digitally re-referenced to the average of the two mastoid electrodes. Data was band-pass filtered between .1 Hz and 100 Hz. Vertical and horizontal eye movement artifacts were identified and removed by subtraction using automatic correlation algorithms within BrainVision Analyzer 2. Other sources of artifact were removed by algorithmically excluding EEG segments with excessive voltage changes (> 50 microvolts per single sample, >150 microvolts per 150 ms) or too little voltage change (< 0.5 microvolts per 100 ms). Individual electrode channels with excessive variance were recovered by interpolation from the surrounding channels. One-second epochs were generated from the continuous EEG starting 100 ms prior to each final word. Individual Event-Related Potential (ERP) averages were

calculated for each condition as specified by the digital trigger codes. These individual participant ERP waveforms were averaged across subjects to make grand average ERPs for each condition for initial visual inspection of results.

3.2.2 Initial ERP Results and Analysis Plan

We used the grand average waveforms from all trials, including both Active/Passive and Reciprocal constructions, to identify and operationalize major peaks; this is an example of the *collapsed localizer* approach which can help reduce false positive findings (Luck & Gaspelin, 2017).

These grand average waveforms were initially split by visual field of presentation; inclusion of the Active/Passive constructions, which constituted 2/3 of the total trials, helped ensure maximum power to assess hemispheric effects. Visual inspection of the grand average waveforms to the lateralized final words revealed an early posterolaterally-maximal positive-negative deflection with maximal amplitudes at about 130 ms and 180 ms, followed by a broadly and centrally distributed slow positive deflection peaking about 380 ms after stimulus presentation. The most positive datapoint from 90–200 ms at each electrode was operationalized as P130, the most negative datapoint from 150–250 ms was operationalized as N180, and the most positive datapoint from 330–485 ms was operationalized as P380.

P130 latency showed a marked effect of visual field of presentation. At the P07-P08 electrode pair, where these deflections attained maximal amplitudes, a 2 x 2 ANOVA (hemisphere x VF) showed a significant interaction, $F(1,19) = 102.51$, $p < .001$, $\eta_p^2 = .84$, as both peaked about 30 ms earlier over contralateral as compared to ipsilateral scalp (Figure 9). This difference has been used in prior studies as a measure of callosal transfer time (Rugg et al.,

1984; Saron & Davidson 1989; Moes et al., 2007), and validates both effective lateralized presentation and ERP recording methods in Experiment 2.

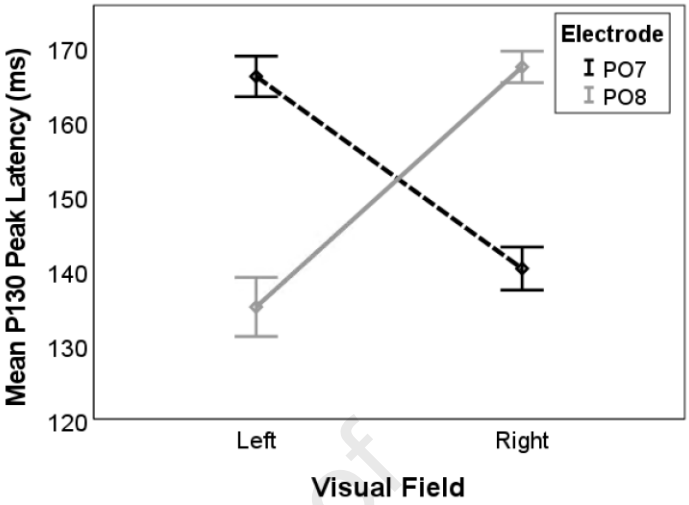

Figure 9. Laterality effects at P130 peak.

An analogous analysis of P380 latency showed a crossover interaction similar to that observed for the P130 peaks, $F(1,19) = 7.30$, $p = .014$, $\eta_p^2 = .28$, though smaller in magnitude, with contralateral presentations producing P380 peaks about 10 ms earlier than ipsilateral presentations. Still, because the linguistic variables of primary interest appeared to alter P380 *amplitude*, but not latency, we collapsed across visual field in subsequent analyses of candidate electrophysiological measures of LE (Φ(LE)). Because the LE behavioral effects were strongest for Reciprocal trials, subsequent ERP analyses were restricted to those trials.

In this exploratory search for an electrophysiological correlate of LE, we used two complementary techniques to quantify effects of the linguistic variables on the centroparietal midline late positive deflection that was both hypothesized (as a P3b effect; Polich, 2007) and visible as a divergence between the A and D grand means that was largest at the expected centroparietal midline electrodes between 300 and 400 ms. The first technique was a temporal bin analysis of the amplitude of five 20 ms temporal bins: (300-320, 320-340, 340-360, 360-380, and 380-400 ms). The advantage of this analysis is that it is completely algorithmic and does not require experimenter interaction to quantify amplitude effects of the experimental variables over the assessed time period of an ERP. The averaged amplitude for each bin from the four

individual ERPs representing the conjunction of Grammatical Subject (1st Person, 3rd Person) and Context (Null, Priming) at three midline centroparietal electrodes (Pz, CPz, Cz) were exported for statistical analysis in SPSS.

The second technique was a peak picker analysis, which assessed latency and amplitude effects on the maximal P380 deflection at the same electrodes. This type of analysis initially selects the data point of maximal amplitude (for a positive deflection) in each condition for each subject at each electrode within the time range specified for the deflection (330 – 485 ms here). Because anomalous calls can result when the deflection of interest is superimposed on excess noise, particularly residual alpha and slow waves, algorithmically selected peaks are visually inspected and corrected to a secondary peak as needed if they are at the first or last data point of the assessed latency range. Eight percent of the automatically identified peaks were corrected to a secondary peak after visual inspection suggested that the initial peak selected was largest due to superimposition of a slow wave or residual alpha activity. The advantages of the peak-picker method are 1) it generates measures of both amplitude and latency, 2) it measures a deflection at individually selected time-points across a wide time range, thereby reducing effects of individual differences in the latency of a peak on the measured amplitude and 3) it allows for superior correction of noise due to slow waves and residual alpha. The disadvantage of this method is that it requires experimenter viewing/correction of some of the algorithmically-selected peaks, which is time-consuming and can potentially produce experimenter confounds. We minimized the probability of the latter by making the experimenter (SB) blind to the condition of the averages viewed.

3.2.3 Linguistic ERP Results

In the initial full analysis, an early frontal component showed significant effects involving the linguistic variables but they were confounded with large residual motor artifacts from EOG (eye movements) and were dropped from further analyses as uninterpretable. We also observed that the positive deflection at 380 ms was sensitive to the linguistic variables as described below.

3.2.3.1 Temporal bin analysis

3.2.3.1.1 ANOVA for Temporal bin analysis

We specifically considered whether the validity criteria for the electrophysiological correlate of LE were satisfied across either the entirety or a subset of the five bins from 300-400 ms. The criteria were (1) a Context x Grammatical Subject interaction or an Electrode (Pz, CPz, Cz) x Context x Grammatical Subject interaction, and (2) Φ(LE) = $\Phi(A) - \Phi(D)$ were significantly different (the "nonzero criterion") and Φ(B) and Φ(C) were between them (the "range criterion").

An initial 4-way ANOVA (Temporal Bin x Electrode x Context x Grammatical Subject) showed that the presence of critical interactions varied by temporal bin, as there was a Bin x Context x Grammatical Subject interaction, $F(4, 76) = 2.97$, $p = .025$, $\eta_p^2 = .14$, and a Bin x Electrode x Context x Grammatical Subject interaction, $F(8, 152) = 2.33$, $p = .022$, $\eta_p^2 = .11$. Planned comparisons showed that 3-way interaction effects within each bin were either significant or marginal ($p < .10$) for the three later bins (340-400 ms), but not for the two earliest bins (300-340 ms). Specifically, the Electrode x Context x Grammatical Subject interaction was significant in the 340-360 ms bin, $F(2, 38) = 3.45$, $p = .042$, $\eta_p^2 = .15$, marginal in the 360-380

ms bin, $F(2, 38) = 2.98$, $p = .063$, $\eta_p^2 = .14$, and significant in the 380-400 ms bin, $F(2, 38) = 3.31$, $p = .047$, $\eta_p^2 = .15$. This effect was not significant in the 300-320 ms or 320-340 ms bins, $F(1, 19) < 1.06$, $p > .36$, $\eta_p^2 < .05$. The Context x Grammatical Subject interaction was not significant in the 340-360 ms bin, $F(1, 19) = 1.86$, $p = .19$, $\eta_p^2 = .09$, was marginal in the 360-380 ms bin, $F(1, 19) = 3.55$, $p = .075$, $\eta_p^2 = .16$, and was marginal in the 380-400 ms bin, $F(1, 19) = 3.88$, $p = .064$, $\eta_p^2 = .17$, with no evidence for this effect in the 300-320 ms or 320-340 ms bins, Fs < 1.

Based on these initial findings, we repeated the 4-way ANOVA while only including data from the 3 later bins. This analysis showed a significant Electrode x Context x Grammatical Subject interaction, $F(2, 38) = 3.41$, $p = .044$, $\eta_p^2 = .15$, as well as a marginal Context x Grammatical Subject interaction, $F(2, 38) = 3.28$, $p = .086$, $\eta_p^2 = .15$. Temporal bin did not interact with either effect, $F < 1$. Thus, for subsequent analyses, we collapse across the time window from 340-400 ms. Since the linguistic variables interacted with electrode, we assessed the validity criteria for each electrode (Cz, CPz, and Pz) separately.

- Cz (Figure 10, left panel): A-D was significantly different from zero $t(19) = -2.62$, $p = .017$, $d = -.59$ (the "nonzero criterion"), and B and C had values between those for A and D (the "range criterion"). A 2 x 2 repeated measures ANOVA showed that the Context x Grammatical Subject interaction was significant ($F(1, 19) = 5.27$, $p = .033$, $\eta_p^2 = .22$), as was the main effect of grammatical subject, $F(1, 19) = 4.54$, $p = .046$, $\eta_p^2 = .19$, such that 3rd person subjects had higher amplitude than 1st person subjects. There was no main effect of Context, $F(1, 19) = 2.85$, $p = .11$, $\eta_p^2 = .13$. $\underline{(A-D)}_{\Phi LE}$ correlated with $(A\text{-}D)_{Rating}$ ($r = -.51$, $p = .021$), and there was also a significant correlation between $\underline{(A-D)}_{\Phi(LE)}$ and $(A\text{-}D)_{RT}$ (r

= .53, p = .016). The correlation between $\underline{(A-D)_{\Phi(LE)}}$ and EQ was not significant (r = .23, p = .33).

- CPz (Figure 10, right panel): A-D was significantly different from zero, $t(19) = -2.57$, p = .019, d = -.57, and the range criterion was satisfied. The Context x Grammatical Subject interaction was marginal ($F(1, 19) = 3.89$, p = .063, η_p^2 = .17), and there was a main effect of Grammatical Subject ($F(1, 19) = 6.07$, p = .023, η_p^2 = .24), but not a main effect of Context, $F(1, 19) = 1.38$, p = .25, η_p^2 = .07. The correlation of $\underline{(A-D)_{\Phi(LE)}}$ with $(A\text{-}D)_{Rating}$ was significant (r = -.48, p = .031). There was also a significant correlation between $\underline{(A-D)_{\Phi(LE)}}$ and $(A\text{-}D)_{RT}$ (r = .56, p = .010). The correlation between $(A-D)_{\Phi(LE)}$ with EQ, was not significant (r = .28, p = .23).

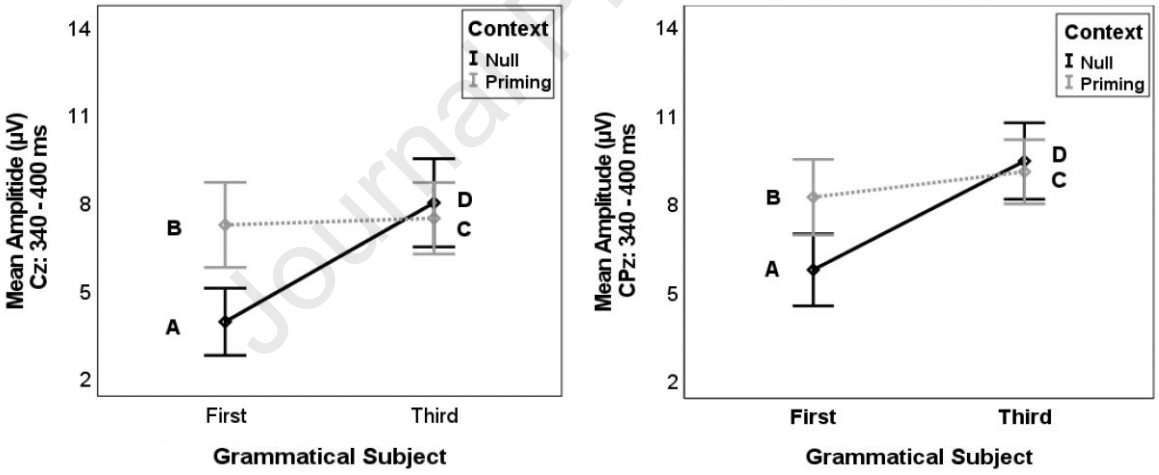

Figure 10. Amplitudes for temporal bin analysis (340-400 ms) at electrodes Cz (left) and CPz (right).

- Pz: A-D was significantly different from zero, $t(19) = -2.52$, p = .021, d = -.56. However, the range criterion was not satisfied, as the peak for C items was slightly higher numerically than the peak for D items, though this difference was not statistically significant, $t(19) < 1$, d = .07. There was also no evidence for a Context x Grammatical Subject interaction, $F < 1$, η_p^2 = .04. There was a main effect of Grammatical Subject, $F(1, 19) = 7.83$, p = .011, η_p^2 = .29,

but not a main effect of Context, $F(1, 19) = 1.96$, $p = .18$, $\eta_p^2 = .09$. The correlation of $\underline{(A-D)_{\Phi(LE)}}$ with $(A\text{-}D)_{\text{Rating}}$ was significant ($r = -.46$, $p = .043$). There was also a significant correlation between $\underline{(A-D)_{\Phi(LE)}}$ and $(A\text{-}D)_{\text{RT}}$ ($r = .55$, $p = .013$). The correlation between $(A-D)_{\Phi(LE)}$ with EQ was not significant ($r = .25$, $p = .28$).

In sum, $(A-D)_{\Phi(\text{LE})}$ for the amplitude range from 340-400 ms appears to provide a good physiological model of the behavioral measure LE, consistent with an empathy hierarchy where A represents no violations and D represents maximum violations of LE. $(A-D)_{\Phi(\text{LE})}$ correlated with LE, defined as $(A\text{-}D)_{\text{Rating}}$, and with the RT difference measure $(A\text{-}D)_{\text{RT}}$, but $(A-D)_{\Phi(\text{LE})}$ did not correlate with EQ.

3.2.3.1.2 Peak-picker analysis

The peak-picker analysis provided estimates for the maximal amplitude and latency of the P380 ERP component for each of the four linguistic conditions (A, B, C, and D), when that peak was allowed to range over the period from 330 to 485 msec independently for each subject at each condition and electrode, rather than integrating the 340-400 sec period as above.[6] We restricted this analysis to electrodes Cz, CPz, and Pz to test our hypothesis that empathy hierarchy violations would produce greater positivity (P3b effect) or negativity (N400 effect) at the centroparietal midline sites most associated with both effects. Because there was no evidence for an N400 peak or N400 effect in grand averages or prior analyses, we did not use the peak-picker method to pick the most negative point.

A 3-way ANOVA (Electrode x Context x Grammatical Subject) on peak amplitude showed evidence for a 3-way interaction, $F(2, 38) = 3.41$, $p = .044$, $\eta_p^2 = .15$, as well as a

[6] The 330-485 ms window was based on grand averages of all trials.

marginal Context x Grammatical Subject interaction, $F(1, 19) = 3.21$, $p = .089$, $\eta_p^2 = .14$. Results for the latency variable at all three electrodes found no evidence for Context x Grammatical Subject or Electrode x Context x Grammatical Subject interactions, Fs < 1. Consequently, we only present results based on the peak-picker amplitude of P380 at Cz, CPz, and Pz electrodes.

- Cz (Figure 11, left panel): $(A - D)_{\Phi(LE)}$ was significantly different from zero ($t(19) = -2.83$, $p = .011$, $d = -.63$) for the peak-picker P380 maximum amplitude at Cz, and the range criterion was met. This measure additionally showed a significant interaction of Context x Grammatical Subject ($F(1, 19) = 5.22$, $p = .034$, $\eta_p^2 = .22$), as well as main effects of Context ($F(1, 19) = 4.56$, $p = .046$, $\eta_p^2 = .19$), and of Grammatical Subject ($F(1, 19) = 6.14$, $p = .023$, $\eta_p^2 = .24$). The main effect of Grammatical Subject indicated that the mean P380 peak for third person subjects was significantly higher than the mean P380 peak for first person subjects. $(A - D)_{\Phi LE}$ correlated with $(A\text{-}D)_{Rating}$ ($r = -.46$, $p = .039$), and there was also a significant correlation between $(A - D)_{\Phi(LE)}$ and $(A\text{-}D)_{RT}$ ($r = .55$, $p = .011$). The correlation between $(A - D)_{\Phi(LE)}$ and EQ was not significant ($r = .27$, $p = .26$).
- CPz (Figure 11, right panel): The analysis at CPz showed largely similar effects. $(A - D)_{\Phi(LE)}$ was significantly different from zero ($t(19) = -2.83$, $p = .011$, $d = -.63$), and satisfied the range criterion. The interaction of Context x Grammatical Subject showed a marginal effect ($F(1, 19) = 4.17$, $p = .055$, $\eta_p^2 = .18$). There was also a main effect of Grammatical Subject ($F(1, 19) = 8.10$, $p = .010$, $\eta_p^2 = .30$), but no main effect of Context, $F(1, 19) = 1.60$, $p = .22$, $\eta_p^2 = .08$. $(A - D)_{\Phi LE}$ correlated with $(A\text{-}D)_{Rating}$ ($r = -.45$, $p = .045$), and there was also a significant correlation between $(A - D)_{\Phi(LE)}$ and $(A\text{-}D)_{RT}$ ($r =$

.61, p = .004). The correlation between $\underline{(A-D)_{\Phi(LE)}}$ and EQ was not significant (r = .29, p = .22).

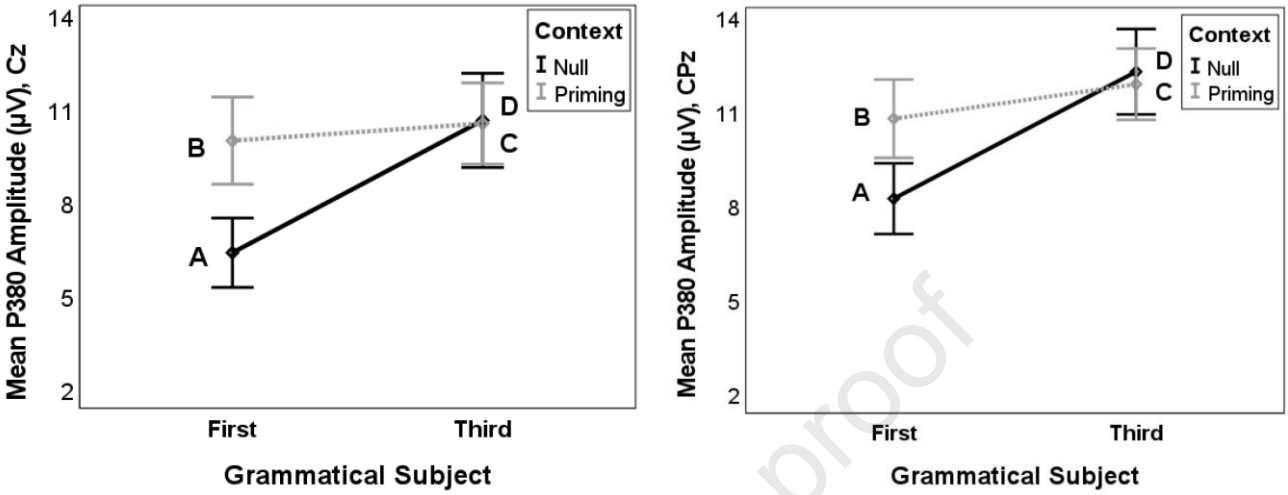

Figure 11. Amplitudes for peak-picker P380 amplitude at electrode Cz (left) and CPz (right).

- Pz: At the Pz electrode, $(A-D)_{\Phi(\mathrm{LE})}$ was significantly different from zero (t(19) = -2.44, p = .025, d = -.55). However, the range criterion was not satisfied, as the peak for C items was numerically higher than the peak for D items; this difference was not significant, however, t(19) < 1, d = .07. There was also no evidence for an interaction of Context x Grammatical Subject, F < 1, η_p^2 = .03. There was a main effect of Grammatical Subject (F(1, 19) = 8.41, p = .009, η_p^2 = .31), but no main effect of Context, F(1, 19) = 1.37, p = .26, η_p^2 = .07. $\underline{(A-D)_{\Phi\mathrm{LE}}}$ was marginally correlated with $(\text{A-D})_{\text{Rating}}$ (r = -.40, p = .084), and there was a significant correlation between $\underline{(A-D)_{\Phi(LE)}}$ and $(\text{A-D})_{\text{RT}}$ (r = .58, p = .007). The correlation between $\underline{(A-D)_{\Phi(LE)}}$ and EQ was not significant (r = .21, p = .39).

Taken together, our results suggest that the Φ(LE) of the peak-picker P380 amplitude is optimal at Cz. It showed expected correlations with $(\text{A-D})_{\text{Rating}}$, and $(\text{A-D})_{\text{RT}}$, suggesting that $(A-D)_{\Phi(\mathrm{LE})}$ shares an information processing mechanism with $(\text{A-D})_{\text{LE}}$, in which violations of

the EH are associated with larger amplitudes of P380 and longer RTs. However, $(A-D)_{\Phi(\mathrm{LE})}$ did not correlate with EQ.

3.2.4 Discussion

The electrophysiological data provided support for the claim that the difference in amplitude of the P380 component elicited by A and D sentences, measured both with the temporal bin method and the peak-picker method, provides a valid electrophysiological correlate of LE at Cz and to a lesser extent at CPz. Thus, Φ(LE) satisfied the following criteria: $(A-D)_{\Phi(\mathrm{LE})}$ was significantly different from zero, ΦB and ΦC had values between those for ΦA and ΦD Φ(LE) showed a significant interaction of Target Subject x Context, and it correlated with the rating measure $(\mathrm{A\text{-}D})_{\mathrm{Rating}}$, as well as with associated RTs, $(\mathrm{A\text{-}D})_{\mathrm{RT}}$.

One remaining question involves whether Φ(LE) shares characteristics with components that have been associated by others with Psychological Empathy. As discussed in the introduction, amplitude of the midline maximal centroparietal P3b, associated with task-related processing of low-probability or unexpected stimuli (Polich, 2007), has indeed been correlated with empathy using pain pictures (Ibanez et al., 2012), with a similar P3-like positivity associated with empathy when discriminating between facial expressions (Choi and Watanuki, 2014). While our Φ(LE) is a complementary P3b effect highly correlated with LE, it did not itself correlate with Psychological Empathy.

In contrast, we found no evidence for an N400 effect like those correlated with Psychological Empathy in van den Brink et al. (2012). While some of our task demands were similar, there were also major differences: (1) their stimuli were presented verbally instead of visually, (2) some of their stimuli contained semantic violations, and their pragmatic violations

dealt with gender and pitch instead of our lexical and structural pragmatic violations of LE, and (3) their participants rated sentences for well-formedness first and then had EEG recorded subsequently instead of simultaneously.

A similar design to ours was used to assess ERP signatures of confirmed and violated semantic predictions (Kuperberg et al., 2019). Participants read three-sentence passages where the final word confirmed or violated two different levels of semantic expectation. As can be seen in their Figure 4, final words that confirmed high constraint expectations (black line) produced a positive peak very similar in spatiotemporal distribution to our P380. Their four levels of semantic violation superimpose an 8 microvolt N400 over this peak, obscuring it completely, whereas our 3 levels of empathy hierarchy violations produce no evidence for an N400, but instead increase the amplitude of the P380, by about 5 microvolts in the double-violation D condition (Figures 10 and 11).

4. GENERAL DISCUSSION

Our study examined four main hypotheses: that (1) Linguistic Empathy (LE) can be defined operationally and measured behaviorally as $(A\text{-}D)_{\text{Rating}} = (\bar{A}_{Rating} - \bar{D}_{Rating})$; (2) LE correlates with Psychological Empathy (PE) (Baron-Cohen & Wheelwright, 2004); (3) There is an electrophysiological correlate of LE (Φ(LE)); and (4) Φ(LE) correlates with PE. The first three hypotheses were supported, but the last hypothesis was not.

Experiment 1 found that the operationalization of LE as $(A\text{-}D)_{\text{Rating}} = (\bar{A}_{Rating} - \bar{D}_{Rating})$ met the criteria to be a valid measure of LE and also found that LE correlates positively with EQ, confirming the first two hypotheses. Experiment 2 repeated Experiment 1 with a new group of participants, who, in addition to ratings, had their RTs and EEG recorded. We again found that

$(A\text{-}D)_{Rating}$ was a valid measure of LE and that it correlated positively with EQ. Turning to the electrophysiological data, we successfully found an electrophysiological model of LE Φ(LE) (confirming the third hypothesis), but Φ(LE) did not correlate with EQ (failing to confirm the fourth hypothesis). We may summarize these data by saying that Φ(LE) is a partial correlate of LE.

4.1 Psychological Aspects of Linguistic Empathy

The following results may shed light on psychological aspects of LE. LE correlated positively with EQ. There were negative correlations between ratings and RTs, both across participants and within participants for D items. Additionally, across participants, EQ correlated positively with the largely negative within-participant correlations of ratings with RTs. The information processing implications of these results are discussed below.

(1) The positive correlation of LE and EQ was primarily driven by people with higher EQ rating D items as less acceptable. Compared to people with lower EQ, they found that sentences violating linguistic EHs sounded worse. This suggests that there is a correlation between greater EQ and sensitivity to violations of EHs, whereas people with lower EQ may rely more on syntactic/semantic grammaticality to rate acceptability.[7] When processing and rating the Linguistic Empathy targets, there may be a general empathic process, perhaps involving perspective taking, which is related to the mental process of Psychological Empathy.

[7] There are some trials that are infelicitous for reasons that are not because of an EH violation (e.g., *I married myself*), but there are too few types of these semantic violations to unequivocally dismiss the possibility that people with higher EQ could have a stronger reaction to all infelicitous sentences.

Some caveats are necessary. It is possible that people with greater EQ tend to have stronger reactions to all infelicitous sentences, whether they are "bad" because of Linguistic Empathy-related constraints, discourse pragmatic constraints generally, or syntactic/semantic violations. Linguistic EHs were the only linguistic variable in these experiments, and the degree to which EQ correlates with sensitivity to other linguistic violations is yet to be tested. It is possible that people with higher EQs are more likely to notice and learn linguistic patterns and therefore recognize violations of those patterns better. Finally, it is possible that both EQ and LE correlate with an unknown third factor.

(2) There were negative correlations between ratings and RTs across participants for A and D items, and within participants for D items. In other words, it took more time for participants to rate items lower. Perhaps participants rate each stimulus by subconsciously comparing it with a small set of schemas of common discourse forms appropriate for the given context. Greater deviation of a stimulus from these common discourse forms could require more processing time. For example, after a Null Context sentence, a sentence with a first-person subject would fit the expected schema, while one with a third-person sentence would not. Yokoyama (1986) established that the mutual awareness by the interlocutors of *{I, you, here, now}* is required for both speakers to engage in conversation, so a third-person entity would not fit into a schema unless it were mentioned or otherwise of mutual concern; to mention a third-person without context would constitute an *imposition*, in which a speaker unilaterally introduces a referent or concept without appropriate context (Yokoyama, 1986, pp. 59 ff., pp. 255 ff.). It would take less time to determine that a stimulus fits a common discourse schema than to seek a schema that it might fit.

(3) Across participants, for D items, EQ correlated positively with the within-participants correlations of ratings and RTs. In other words, people with lower EQs took longer to rate D items lower; those with higher EQs were able to process sentences with EH violations more quickly. It appears that the participants with higher EQ are more adept at recognizing stimuli with multiple EH violations as infelicitous, and thus do so more quickly. With respect to the schema model proposed above, people with higher EQ may be better able to recognize and process these schemas. Participants with lower EQ may recognize that a stimulus appears infelicitous, but could be slower to respond for a couple of reasons, which may be independent or combined: (1) Lower EQ causes them to deploy the schemas slowly or (2) The incongruity that a stimulus sounds unnatural but has no grammatical error is harder to reconcile in lower EQ individuals. The finding that participants with lower EQ generally rated D sentences higher provides additional support for these explanations; when they did rate D items lower, these participants did so even more slowly than did participants with higher EQ, but often D items were rated higher and processed quickly, perhaps because no grammatical error was found.

4.2 Neurophysiological Aspects of LE

The third hypothesis was that there exists an event-related EEG correlate of LE, Φ(LE). In Experiment 2, we found a good candidate: amplitude of the P380, a large positive voltage deflection peaking about 380 ms after presentation of the final word which elicited the acceptability ratings. This deflection was measured at both the Cz and CPz midline electrodes over the 60 ms period of maximum grand mean amplitude and also by using the peak-picker method to find the maximum amplitude in each individual subject at each electrode. P380 met all of our criteria for Φ(LE), except that it was not correlated with EQ. Possible explanations for the

lack of a correlation include low sample size or indirect relationships between LE, PE, and P380 amplitude. For instance, LE, measured through the acceptability ratings, may be correlated with EQ primarily through the emotional factors in the EQ questionnaire, but it may be correlated with P380 amplitude through the cognitive information-processing demands of the final word stimuli that generate both the acceptability ratings and the P380.

4.3 Additional Findings, Limitations, and Suggestions

Absence of laterality effects. We expected a hemispheric difference in LE so we lateralized both the input (visual hemifield: LVF, RVF) and the output (response hand: Lh, Rh) of the context-target sentence pairs in order to compare the two hemispheres (LVF-Lh = right hemisphere; RVF-Rh = left hemisphere). We felt that as a linguistic concept, Linguistic Empathy might be specialized to the left hemisphere, but to the extent that Linguistic Empathy is related to Psychological Empathy, and thus part of natural language pragmatics, it might be more specialized to the right hemisphere. However, we found no interactions involving Hemisphere and the linguistic variables in either experiment. This suggests similar information processing of LE in the two hemispheres.

*Absence of Significant Sex Differences in EQ***.** Previous studies examining sex differences in empathy have produced varied results (Baez et al., 2017). Women have been found to show greater empathy based on self-assessment, but not on experimental or physiological measures. Our lack of sex differences fits the latter pattern but may also be due to small sample sizes (Experiment 1: 19 female, 13 male; Experiment 2: 14 female, 6 male), since previously noted sex differences had small effect sizes, which would only be significant with much larger samples (Baez et al., 2017).

Relatively high ratings for D items, especially among some participants with lower EQs. Though on average, as predicted, the ratings for A items were significantly higher than those for D items, some participants, especially those with lower EQs, rated D items high. Since all sentences contained no semantic or syntactic violations, it seems likely that some participants recognized that all sentences were grammatically valid and thus gave them high ratings despite EH violations.

The use of EQ as a measure of PE. The EQ has some limitations as a measure of Psychological Empathy. Popular self-report inventories of PE consistently reveal separate factors for emotional empathy (bottom-up, possibly mirror neuron based) and cognitive empathy (top-down, inferential) (Batchelder et al., 2017). Unfortunately, our experiment was unable to examine the factorial structure of EQ because the computerized version of the EQ administered to our participants did not provide scores for individual items. However, we can make predictions for follow-up studies. Factor analyses of the EQ have yielded three factors: (1) *cognitive empathy,* (2) *emotional reactivity*, and (3) *social skills* (Baron-Cohen & Wheelwright, 2004; Lawrence et al., 2004; Muncer & Ling, 2006. When Kuno and Kaburaki (1977) first discussed Linguistic Empathy, the emotional aspects of empathy were not involved in the processing or production of EHs. Rather, cognitive understanding of context and relationships was necessary. In this view, rating Linguistic Empathy stimuli would not seem to involve emotional arousal, but rather appear to involve the kind of top-down inferences needed for the ability to attribute mental states to other people, to interpret, explain, and predict their behavior, similar to Theory of Mind. Thus, we would predict that Linguistic Empathy would more strongly correlate with the *cognitive empathy* factor of the EQ than with the *emotional empathy* or *emotional reactivity* factors. In addition to breaking the EQ down into factors, it would be useful

to see if other measures of Psychological Empathy such as the Interpersonal Reactivity Index (Davis, 1980) and the Empathy Components Questionnaire (Batchelder et al., 2017) also correlate with measures of Linguistic Empathy and can shed light on the factors involved.

4.4 Implications and Limitations

As discussed above, we were able to provide a preliminary validation of the proposed operationalization of Linguistic Empathy as $(A\text{-}D)_{Rating}$ and to find that it correlated with EQ. As far as we know, this is the first attempt to define, quantify, and validate a measure of Linguistic Empathy. However, our operationalization of Linguistic Empathy was only partial. Firstly, the Linguistic Empathy measured here applies to language comprehension; the relation of language production to Linguistic Empathy was not examined. Additionally, we did not include violations involving inanimate entities as subjects. There are several lists of other EHs that have been proposed and tested (e.g., Descriptor Empathy Hierarchy, Surface Structure Empathy Hierarchy), and there are other linguistic phenomena related to empathy (e.g., agency, animacy, the mutual awareness by the interlocutors of *{I, you, here, now}*, Personal Empathy) that can potentially be designed as variables for similar experiments. Finally, we did not include Active/Passive constructions in the analyses, which is another means to manipulate some EHs. Kann (2017, pp. 84 ff.) demonstrated significant results for Active/Passive stimuli, most relevantly a positive correlation between (A-D) ratings and EQ in English. As discussed earlier, unlike the Reciprocal stimuli, passive sentences contain syntactic and lexical variation that can potentially confound results, and they were thus omitted from analysis in this study.

Another linguistic contribution from these experiments is to enlarge the scope of discourse pragmatics considered in language experiments. Synonymy is typically operationalized

by the truth conditions of semantics; however, this experiment demonstrated that sentences that contain identical logical truth conditions and even identical references but differ in certain other discourse pragmatic conditions can differ in acceptability and are thus not effectively synonymous. Indeed, this position was predicted and assumed in designing our experiments, and the significant results confirm this assumption.

4.5 Extensions and Applications of the LEAT

One criticism of current PE measures is that these tests are self-report questionnaires that ask direct questions about a person's prosocial emotional behavior. This sort of questioning creates a potential conflict of interest with respect to self-incrimination, social desirability, and emotional contamination. We contend that assessing empathy without asking the participant to self-report on socio-emotional issues would be preferred. As a result of the correlation between LE and PE, a metric like the LEAT that is developed to screen for low EQ would do so without the inherent social desirability conflict such as admitting to negative socio-emotional tendencies.

As mentioned, the preliminary version of the LEAT in its current form investigates only two specific EHs. A next step for the LEAT would contain a different range of EHs among the stimuli so that a broader set of Linguistic Empathy violations could be verified and incorporated. If subsequent experiments maintain a significant correlation of Linguistic Empathy with EQ and other measures of PE, they may be useful in therapeutic or educational settings. Linguistic stimuli could be developed or integrated into therapeutic tools for first language development, second language learning, social-emotional awareness (specifically regarding context), perspective awareness and discourse pragmatics.

Additionally, instead of the LEAT using acceptability ratings, an implicit measure, such as reaction time, is more likely to represent participants' automatic, less self-conscious judgements closer to one's true self-concept, without self-awareness, public identity, or group affiliations. The use of auditory instead of visual stimuli would provide even more precise RTs for language processing as well as the opportunity to manipulate other discourse pragmatic variables, including phonology, prosody, and active/passive constructions. The benefit of an implicit, impartial test that provides a correlating measure with PE through a linguistic channel could sidestep any socio-emotional stigma associated with low Psychological Empathy. Since the notion of an individual measure of Linguistic Empathy is new and not associated with social desirability, Linguistic Empathy would be comparatively less stigmatized. Although this study cannot directly address the unfortunate social stigma associated with low PE, it is our hope that the stigma-free notion of the individual measure of Linguistic Empathy can help to normalize the discourse around varying levels of empathy.

4.6 Ultimate Questions

In both experiments, our analysis revealed a strong positive correlation between LE and PE, and between Φ(LE) and LE, consistent with the view that Linguistic Empathy involves a real-time process that parallels perspective taking in Psychological Empathy. The reason for the correlation of PE and LE may be the existence of some general ability that affects the responses both to the PE questionnaire and our stimuli.

The significant correlation of EQ with within-individual RatingRT correlations suggests a process that unfolds in real time so that the distinction between linguistic structure and psychological function is blurred. A natural question is whether the relationship of EQ to the

RatingRT correlation is informative regarding the formation/evolution of LE. Linguistic Empathy, evidently, addresses our underlying communicative needs whereby it gets encoded by a speaker and decoded by an addressee (with sufficiently high EQ) to their mutual communicative satisfaction. The fact that Psychological Empathy matters, as this paper shows, suggests that empathy-based linguistic choices pertaining to subject choice (in English), context, and the choice of referential expressions are acquired to different degrees by those with higher/lower EQ. Linguistically, developmental factors and diachronic changes are likely to correlate. Thus, with the importance of empathy in language use, the possibility that the correlation describes a phylogenetic or an ontogenetic process is tempting. The choices are there in the linguistic structure, and as this paper shows, those with lower EQ somehow may not be as efficient in accessing these structures that satisfy empathy-related communicative needs; perhaps these needs are of less concern for them, or perhaps they are able to satisfy these needs using other communicative tactics. It is also possible that the presumed "real-time processing effects" demonstrated for Linguistic Empathy in these experiments do not reflect adaptive functional phenomena, either in real time or in evolution, and are side effects of other processes or may even be merely vestigial.

References

Baez, S., Flichtentrei, D., Prats, M., Mastandueno, R., García, A. M., Cetkovich, M., et al. (2017). Men, women … who cares? A population-based study on sex differences and gender roles in empathy and moral cognition. *PLoS ONE*, *12*(6), e0179336. https://doi.org/10.1371/journal.pone.0179336

Baron-Cohen, S. & Wheelwright, S. (2004). The empathy quotient: An investigation of adults with Asperger Syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders, 34*(2), 163-175.

Batchelder L., Brosnan M., & Ashwin C. (2017). The development and validation of the empathy components questionnaire (ECQ). *PLoS ONE*, *12*(1), e0169185. https://doi.org/10.1371/journal.pone.0169185

Bryant, B., (1982). An index of empathy for children and adolescents. *Child Development*. *53*(2), 413-425.

Choi, D., & Watanuki, S. (2014). Effect of empathy trait on attention to faces: An event-related potential (ERP) study. *Journal of Physiological Anthropology*, *33*(4). doi:10.1186%2F1880-6805-33-4.

Deane, P. (1992). *Grammar in mind and brain: Explorations in cognitive syntax.* Berlin: Mouton de Gruyter.

Davis, M. H. (1980). *Interpersonal Reactivity Index (IRI)* [Database record]. APA PsycTests. https://doi.org/10.1037/t01093-000

Ferguson, H. J., Cane, J. E., Douchkov, M., & Wright, D. (2015). Empathy predicts false belief reasoning ability: evidence from the N400. *Social Cognitive and Affective Neuroscience*, *10*(6), 848–55.

Hecht, D. (2014). Cerebral lateralization of pro- and anti-social tendencies. *Experimental neurobiology*, *23*(1), 1-27. https://doi.org/10.5607/en.2014.23.1.1

Hogan, R. (1969). Development of an empathy scale. *Journal of Consulting and Clinical Psychology*, *33*(3), 307–316.

Ibanez, A., Melloni, M., Huepe, D., Helgiu, E., Rivera-Rei, A., Canales-Johnson, A., & Moya, A. (2012). What event-related potentials (ERPs) bring to social neuroscience? *Social Neuroscience, 7*(6), 632-649.

Jolliffe, D., & Farrington, D. P. (2006). Development and validation of the basic empathy scale. *Journal of Adolescence*, *29*(4), 589-611.

Kann, T. (2017). *Measuring linguistic empathy: An experimental approach to connecting linguistic and social psychological notions of empathy.* [Doctoral dissertation, UCLA]. https://escholarship.org/uc/item/3zt4r833.

Kohler, W. (1929). *Gestalt Psychology*. New York, NY: Liveright.

Kuno, S. & Kaburaki, E. (1977). Empathy and syntax. *Linguistic Inquiry, 8*(4): 627-672.

Kuno, S. (1987). *Functional syntax: Anaphora, discourse, and empathy.* The University of Chicago Press.

Kuperberg, G. R., Brothers, T., & Wlotko, E. W. (2019) A tale of two positivities and the N400: Distinct neural signatures are evoked by confirmed and violated predictions at different levels of representation. *Journal of Cognitive Neuroscience*, *32*(1), 12-35.

Kutas, M., & Hillyard, S. A. (1980). Event-related brain potentials to semantically inappropriate and surprisingly large words. *Biological psychology, 11*(2), 99-116.

Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating N400 effects of prediction from association in single-word contexts. *Journal of Cognitive Neuroscience*, *25*(3), 484-

502.

Lawrence, E. J., Shaw, P., Baker, D., Baron-Cohen, S., & David, A. S. (2004). Measuring empathy: Reliability and validity of the Empathy Quotient. *Psychological Medicine, 34*(5), 911–920.

Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology,* 54, 146–157.

Mead, G. H. (1934). *Mind, self, and society.* Chicago, IL: University of Chicago Press.

Moes, P. E., Brown, W. S., & Minnema, M. T. (2007). Individual differences in interhemispheric transfer time (IHTT) as measured by event related potentials. *Neuropsychologia*, *45*(11), 2626-2630.

Muncer, S. J., & Ling, J. (2006). Psychometric analysis of the empathy quotient (EQ) scale. *Personality and Individual Differences, 40*(6), 1111-1119.

Oshima, D. Y. (2006). *Perspectives in reported discourse* [Doctoral dissertation, Stanford University]. https://www.gsid.nagoya-u.ac.jp/oshima/docs/dissertation_filed_ver.pdf.

Oshima, D. Y. (2007). Syntactic direction and obviation as empathy-based phenomena: a typological approach. *Linguistics: An Interdisciplinary Journal of the Language Sciences, 45*(4), 727–763.

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clinical Neurophysiology, 118*(10), 2128-2148.

Rugg, M. D., Lines, C. R., & Milner, A. D. (1984). Visual evoked potentials to lateralized visual stimuli and the measurement of interhemispheric transmission time. *Neuropsychologia*, *22*(2), 215-225.

Saron, C. D., & Davidson, R. J. (1989). Visual evoked potential measures of interhemispheric

transfer time in humans. *Behavioral Neuroscience, 103*(5), 1115–1138.

Shamay-Tsoory, S. G., Aharon-Peretz, J., & Perry, D. (2009). Two systems for empathy: A double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions. *Brain*, 132(3), 617-627.

Silverstein, M. (1981). Case marking and the nature of language. *Australian Journal of Linguistics, 1*(2), 227-244. DOI 10.1080/07268608108599275

Silverstein, M. (2016). Hierarchy of features and ergativity. In P. Muysken & H. van Riemskijk (Eds.), *Features and projections* (pp. 163-232). Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110871661-008.

Taylor, I. & Taylor, M. M. (1990). *Psycholinguistics: Learning and using language*. Englewood Cliffs, NJ: Prentice Hall.

Toepoel, V., Vermeeren, B., & Metin, B. (2019). Smileys, stars, hearts, buttons, tiles or grids: Influence of response format on substantive response, questionnaire experience and response time. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique, 142*(1), 57–74. https://doi.org/10.1177/0759106319834665

van den Brink, D., van Berkum, J. J., Bastiaansen, M. C., Tesink, C. M., Kos, M., et al. (2012). Empathy matters: ERP evidence for inter-individual differences in social language processing. *Social Cognitive Affective Neuroscience, 7*(2), 173-183. PMCID: PMC3277364

Yokoyama, O.T. (1986). *Discourse and word order.* Philadelphia: John Benjamin.

Yokoyama, O.T. (1999). The speech act empathy hierarchy and Russian possessives. In K. I. Takami & A. Kamio (Eds.), *Function and Structure* (pp. 57-82). Philadelphia: Benjamins.

Yokoyama, O.T. (2000). Èmpatija v ramkax transkacionnoj modeli diskursa [Empathy in transactional discourse model]. In L.P. Krysin (Ed.), *Russkij jazyk segodnja* (pp. 276-286). Moscow: Azbukovnik.

Yokoyama, O.T. (2019). Ot zerkal'nyx nejronov k reči: èmpatija v mozgu i v jazyke [From mirror neurons to speech: Empathy in the brain and in language]. *Jazyk i metod*, *6*, 111-124.

Zaidel, E. (1998). Language in the right hemisphere following callosal disconnection. In B. Stemmer & H. A. Whitaker (Eds.), *Handbook of neurolinguistics* (pp. 369-383). Academic Press.

Appendix

Background Questions:

1. Name (Or Participant #) _____

2. Age: ( ) Under 18 ( ) 18-25 ( ) 26-35 ( ) 36-45 ( ) 46-59 ( ) 60+

3. Experience with English:
   ___ Native Fluency: ( ) Monolingual Native ( ) Multi-lingual Native
   ___ Non-native fluency: Fluency began ( ) under 10-years-old ( ) 10-18 ( ) 18+
   ___ Non-native proficient: Studied for ( ) 0-2 years ( ) 3-5 years ( ) 6+ years

Briefly describe your experience with English:

_____

_____

4. Cultural Background
   a. Born and raised in California: ( )Y or ( )N
   b. Born and raised in USA: ( )Y or ( )N
      i. If "no" to a-c where were you born? _____
      Raised?_____
   c. Born abroad and raised in USA: ( )Y or ( )N
   d. One or both parents born and raised outside California: ( )Y or ( )N
   e. One or both parents born and raised outside USA: ( )Y or ( )N
      i. If "yes" to d-e where were they born?_____
      Raised?_____
   f. What languages do your parents speak?

_____

   g. First names of your immediate family?

_____

   h. How do you self-identify? Check all that apply.
      i. American ____
      ii. Black/African-American ____ non-US African ____
      iii. Latino/a/x ____ non-US Latin-American ____
      iv. Asian-American ____ non-US Asian ____
      v. White/Caucasian-American ____ non-US white/caucasian ____
      vi. Middle-East-American ____ non-US Middle-Eastern ____
      vii. Central-Asian-American ____ non-US Central-Asian ____
      viii. Multiethnic American ____ non-US multiethnic ____
      ix. Other (please specify)

      _____

5. Life as a student:
   How many years at a university have your completed?
   ( ) 0 ( ) 1 ( ) 2 ( ) 3 ( ) 4 ( ) Graduate Student
   What is your major? Minor?

   _____

   What, if any, do you consider your other areas of expertise?

   _____

Highlights

- Preliminary measure of Linguistic Empathy correlated with Psychological Empathy
- Linguistic Empathy shares information processing with Psychological Empathy
- Linguistic Empathy has a partial electrophysiological correlate
- Logically synonymous phrases that differ pragmatically are not fully synonymous