# Utilizing Coordinated Data Analysis to Examine Structural, Linguistic, and Discursive Predictors of Thought-Feeling Inferences in Inferential Accuracy Research

by

Zachary J. Schroeder

A dissertation accepted and approved in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in Psychology

Dissertation Committee:

Sara D. Hodges, Chair

Vishnu “Deepu” Murty, Core Member

Sara J. Weston, Core Member

Jenefer Husman, Institutional Representative

University of Oregon

Spring 2026

DISSERTATION ABSTRACT

Zachary J. Schroeder

Doctor of Philosophy in Psychology

Title: Utilizing Coordinated Data Analysis to Examine Structural, Linguistic, and Discursive Predictors of Thought-Feeling Inferences in Inferential Accuracy Research

When perceivers attempt to infer what another person is thinking and feeling during a conversation, the most robust predictor of their accuracy is the verbal information available in the interaction. Yet prior research in this area (inferential accuracy) has treated verbal information as a monolithic construct, asking only whether its presence, absence, or intelligibility affects accuracy rather than examining which specific features of speech predict perceiver inferences and target thoughts. In three studies, I used sentence embeddings to compute the semantic similarity between transcribed speech and either perceiver inferences (cue utilization) or target-reported thoughts (cue validity). Data from six previous inferential accuracy studies were synthesized using a data analysis framework combining Bayesian multilevel Gaussian process models with individual participant data meta-analyses.

In Study 1, although perceiver inferences were not more similar to target speech than the target's interactant's speech as hypothesized, they were more semantically similar to speech occurring near the end of a video chapter, an effect driven primarily by interactant speech in standard stimulus (interview-style) paradigm studies. Study 2 extended these findings to the level of dialogue acts (e.g., inform, question, suggestion, agreement), revealing that inform and question dialogue acts showed unique temporally distributed patterns of alignment with perceiver inferences, but again only in standard stimulus paradigm studies where the interactant served as an interviewer. Study 3 replaced perceiver inferences with target-reported thoughts,

seeking to identify which findings from the correlates of cue utility actually mapped onto cue validity. Target thoughts were also most similar to recent speech, and no speaker difference was observed, but the most consequential discursive finding was that suggestion dialogue acts showed the strongest alignment with target thought content of any dialogue act category, identifying a possible point of miscalibration.

Together, these findings provide a preliminary map of both sides of the lens model for verbal information in inferential accuracy. The structural dimensions of the cue space (temporal position, speaker identity) show calibration between utilization and validity: perceivers' inferences are most similar to recent speech, and recent speech is indeed most similar to what targets were thinking. The discursive dimensions reveal a specific misalignment: the dialogue act category most semantically similar to target thought content (suggestions) is not correspondingly reflected in perceiver inferences, identifying a potential target for improving perceiver accuracy. The embedding-based similarity measure, Gaussian process temporal models, and meta-analytic trajectory synthesis developed here provide a framework for extending this map to new conversational contexts and finer-grained dialogue act taxonomies.

ACKNOWLEDGMENTS

I would like to begin by thanking my mentors, and first on that list is my advisor, Sara Hodges, whose years of patient guidance made me a better researcher than I ever could have been without her. I would also like to express my deepest gratitude for my committee: Deepu Murty, Sara Weston, and Jenefer Husman. Their advice and input made this project infinitely better. I would like to thank the members of the Social Cognition lab, Erika Davis and Eliott Doyle, and give special thanks to Murat Kezer, whose research integrating NLP with inferential accuracy inspired this dissertation. To Lori Olsen, your compassion and wisdom has been the compass guiding me through every step of this process. I also want to thank Wendi Gardner and Sara Broaders, whose mentorship to an eager and annoying undergraduate back at Northwestern University made me realize that pursuing a PhD was something that I could actually accomplish. To A.P. Bevels and D.L. Cartwright, your work has inspired me from afar and I have tried my best to emulate your clarity in storytelling, wit, and mastery of the English language here.

I want to also thank my family and friends whose support made all of this possible. Mom, Dad, Ben, and Louise, thank you for everything. To my grandparents, thank you for everything you have taught me over the last thirty years. To better friends than I could have ever imagined meeting when I moved to Oregon, Raleigh, Ann-Marie, and Kavya; to those far away who dealt with me calling during the workday, Izzy and Maggie; and to the Sprinklers of Wayne, who made Thursday nights the highlight of many very long weeks. And to Cody, who entered my life at my most frazzled and who daily makes me want to be a better person, thank you.

This research was supported by a National Science Foundation Graduate Research Fellowship (NSF-GRFP) grant number 2236419.

For Goose.

TABLE OF CONTENTS

Chapter Page

LIST OF FIGURES

LIST OF TABLES

# I. INTRODUCTION

Of the many strange phenomena in the social media era, the proliferation of "body language experts" on video-sharing sites like YouTube may be of particular fascination to a researcher studying inferential accuracy (the accurate inference of the thoughts and feelings of another person; Gesn & Ickes, 1999; Hodges & Kezer, 2021; Ickes, 2001; Ickes et al., 1990; Levenson & Ruef, 1992). In the days after a celebrity scandal, bomb-shell interview, or viral moment, these "experts" post hundreds of videos analyzing the body language of the celebrity in question, and in turn these videos can accrue millions of views. For example, one video by "The Behavior Panel" (2021) analyzing Prince Harry and Meghan Markle's body language in their tell-all interview with Oprah Winfrey amassed over 4.7 million views – at just under two hours long, that's a combined 1,055 years of watch-time of people wanting to find the hidden meaning "under the surface" of the interview (Winfrey, 2021).

However, without knowing what Meghan and Harry were actually thinking in that interview, these viewers were buying into a pseudo-scientific mishmash about "body language" (for a review, see Patterson et al., 2023). In contrast, inferential accuracy research has accumulated over 40 years of empirical evidence, though less sensational than some popular interpretations, about people's ability to accurately perceive what other people are thinking and feeling. Perceivers are generally more accurate than would be expected by pure chance when they are shown a video of a social interaction and asked to infer what a target person is thinking and feeling (Ickes, 2011). However, the traits and skills that some body language experts purport to be high in such as nonverbal decoding ability or self-reported interpersonal sensitivity seem to play a relatively small role in accurately inferring someone's thoughts and feelings (Stinson & Ickes, 1992). Although it may be unsatisfying to these body language experts and the people who

watch their videos, inferential accuracy research has pointed to using what someone says as the most robust cue to what they are thinking and feeling (Gesn & Ickes, 1999; Hall & Schmid Mast, 2007; Hodges et al., 2010; Hodges & Kezer, 2021; Kraus, 2017).

Thus, this dissertation embraces the importance of words in inferential accuracy and the complexity that comes with them. In doing so, I depart from the question that has organized most prior inferential accuracy research (who is a good "mind-reader"?) and ask instead how perceivers use verbal information to arrive at their inferences in the first place. Prior work has asked variations on the question 'Does X predict inferential accuracy?' Across studies, X has been replaced with a slew of constructs (reviewed below), such as gender of the perceiver (or target), perceiver's motivation, and perceiver's familiarity with the target. Of all predictors of accuracy, availability of and use of the target's verbal information has emerged as one of the most, if not *the* most, consistent.

I build this investigation on a single fundamental research question: setting accuracy aside as the outcome of interest, can the aspects of verbal information that perceivers attend to when making thought-feeling inferences be identified and modeled? This chapter (Chapter I) provides a background of inferential accuracy research, describes the two most prominent research paradigms, grounds the investigation within Brunswik's (1956) lens model, and reviews the evidence for predictors of inferential accuracy. It ends by describing the rationale for focusing on verbal information as the key to understanding how perceivers infer targets' thoughts and feelings.

Chapter II begins with a detailed description of the analytical methods used in Studies 1, 2, and 3. These include the natural language processing tools used to transform transcribed social interactions into quantitative data as well as a review of the Bayesian multilevel modeling

approach used for individual participant data meta-analyses, which pooled findings across inferential accuracy studies. It then introduces data from six past inferential accuracy studies, examining structural aspects of inferential accuracy stimuli that may predict which verbal information is particularly similar to perceivers' thought-feeling inferences (e.g., whether verbal information may be weighted differently depending on when it occurs during a stimulus video).

Chapter III presents Study 2, which examines verbal information at the level of dialogue acts. In it, I categorize all utterances within each social interaction from the previous studies into one of seven dialogue act categories (*acknowledgment*, *agreement*, *disagreement*, *functional*, *inform*, *question*, and *suggestion*) using the Dialogue Act Markup in Several Layers (DAMSL; Core & Allen, 1997) coding scheme. Chapter IV returns to an investigation of accuracy, with Study 3 which extends the findings from Studies 1 and 2 to examine which aspects of verbal information correspond to targets' thoughts and feelings (i.e., which cues carry the most validity as indicators of target thought content). Finally, Chapter V concludes the dissertation with a general discussion synthesizing the findings and implications of all three studies.

**Inferential Accuracy**

Researchers studying inferential accuracy seek to understand the ways a perceiver may accurately infer the thoughts and feelings of a target (Gesn & Ickes, 1999; Hodges & Kezer, 2021; Ickes, 2001; Ickes et al., 1990; Levenson & Ruef, 1992). Although inferential accuracy has previously been referred to as empathic accuracy (e.g., Ickes et al., 1990), recent work by leaders in the field (Davis et al., 2026) has suggested changing the name to "inferential accuracy" as the "empathic" label inaccurately suggests that such accuracy will be accompanied by prosocial

motivations or outcomes[1] (Hodges et al., 2024). As such, I use the term 'inferential accuracy' to describe a perceiver's accuracy for inferring a target's thoughts and feelings.

First described within the clinical psychological research literature, the construct that would become inferential accuracy was initially labeled "accurate empathy" and conceptualized as a clinician's ability to infer a client's emotions and verbalize their inference accurately (Rogers, 1957; Truax & Carkhuff, 1967). This "accurate empathy" was operationalized as a score on the Accurate Empathy Scale, in which a panel of experts (e.g., psychiatrists, clinicians, clinical graduate students) used this scale to rate how "with" the client a therapist was during therapist-client interactions (Truax, 1961). For example, on the nine-point scale, the lowest accuracy score of one was labeled, "therapist seems completely unaware of even the most conspicuous of the client's feelings" (Truax, 1972, p. 398). However, subsequent research suggested that this method of measuring accurate empathy – i.e., relying on judges' ratings without reports from clients on what they were actually feeling in that moment – was confounded by the judges' holistic view of how good the therapist was generally and not related to the therapist's accuracy in that specific moment (Chinsky & Rappaport, 1970; Kiesler et al., 1967; Rappaport & Chinsky, 1972).

Accurate empathy developed simultaneously with a larger field of interpersonal accuracy, which, by the 1980s, consisted of three primary lines of research: accuracy in judging personality traits (e.g., Funder & Colvin, 1988), accuracy in meta-perceptions (i.e., accuracy in how one is perceived by others; Sillars & Scott, 1983), and accuracy in emotion perception (e.g., Hall, 1978). Building off of these, William Ickes developed his conceptualization of "empathic

[1] For example, a poker player may be a highly accurate perceiver but not prosocial in taking their opponent's money.

accuracy" as accuracy for inferring discrete thoughts and feelings during a dynamic, dyadic social interaction which he measured using what he dubbed the *dyadic interaction paradigm* (Ickes, 1982, 1993; Ickes et al., 1990; Marangoni et al., 1995).

In the dyadic interaction paradigm, two participants engage in a video-recorded verbal interaction. Immediately following the interaction, one participant (the target) watches the video and marks where they remember having had discrete thoughts and feelings. The other participant, the perceiver, then watches the interaction video and, at the times when the target marked their thoughts and feelings, the video is paused and the perceiver is asked to infer what the target was thinking and feeling at that moment (Gesn & Ickes, 1999; Ickes, 2016; Ickes et al., 1990). These thoughts and feelings – reported by the target and inferred by the perceiver – are then rated for accuracy on a 4-point[2] scale by a team of coders. The scale ranges from 0 (essentially different content) to 3 (essentially the same content; Hodges et al., 2015; Ickes et al., 1990).

The dyadic interaction paradigm was later joined by the *standard stimulus paradigm*, also developed by Ickes (e.g., Marangoni et al., 1995). In it, targets are videotaped in an interaction with either a researcher or conversation interactant[3] (sometimes called the interrogator – a term borrowed from Kagan, 1977). Targets again provide their thoughts and feelings upon rewatching the video of their interaction, but perceivers in this paradigm do not interact with the target during the course of the research process and are often recruited after the video stimuli are created (e.g., Hodges & Kezer, 2021). A similar research paradigm was developed by Kagan

[2] Inferential accuracy studies prior to 2014 used a 3-point scale to code inferential accuracy. Lewis (2014) made the case to change this to a 4-point scale to increase the variability in accuracy scores, and many subsequent inferential accuracy studies followed suit.
[3] In dyadic interaction paradigm studies, the perceiver and interactant are the same person, whereas they are separate people in the standard stimulus paradigm.

(1977) for use in clinical settings and asked perceivers to choose the content of the target's thoughts from a series of multiple-choice options. However, the Ickes standard stimulus paradigm was developed for use in dyadic interactions more generally (not just clinical sessions, although the Ickes paradigm, with perceivers generating inferences in response to open-ended questions, has been used in simulated clinical sessions – see Marangoni et al., 1995).

Although the standard stimulus and dyadic interaction paradigm are the focus of this investigation and are common operationalizations of inferential accuracy, a few other paradigms merit mention. First, adapting earlier methods for studying marital conflict (e.g., Levenson & Gottman, 1983), Levenson and Ruef (1992) videotaped married couples in conversation about a conflict in their relationship. After the short (15-minute) interaction, participants watched the videotape of their interaction with their spouse and manipulated a joystick called a "rating dial" (Levenson & Ruef, 1992, p. 237) to continuously measure their own (target's) affect during the discussion. This affective data (ranging from very negative to very positive affect) was then compared to data from their spouse (the perceiver) who, when watching the same videotaped interaction, attempted to accurately rate the target's affect as it changed over the course of the interaction using the same device. Inferential accuracy in this case was calculated by comparing the time-series correlations of the affect rating dial scores between target and perceiver over the course of the interaction.

Less common than the dyadic interaction and standard stimulus paradigm studies (which measure accuracy for thought/feeling content), or the continuous rating dial paradigm (which measures accuracy for affect valence and intensity), some researchers have collected Likert scale ratings of discrete emotions from targets and perceivers at set intervals during the playback of a recorded interaction as an operationalization of inferential accuracy (e.g., Kraus, 2017;

Kunzmann et al., 2018), or asked targets and perceivers to provide holistic ratings of general positive and negative affect for the whole video stimulus (Ma-Kellams & Blascovich, 2013). In yet another variation, targets' open-ended responses about what they were thinking and feeling were converted into short multiple-choice options from which perceivers were asked to choose the correct inference (e.g., Fernandez-Duque et al., 2010).

### *Predictors of Inferential Accuracy*

As alluded to earlier, few individual differences in perceivers have accumulated robust empirical support for predicting inferential accuracy. This section reviews the literature associated with predictors of inferential accuracy, ending with the predictor of interest in this dissertation, the utilization of verbal information.

**Gender.** The majority of Americans believe that women are more empathic than men (Martingano et al., 2025), a belief that is supported by a consistent gender difference in self-reported empathy (e.g., women generally score higher than men on all four subscales of the Interpersonal Reactivity Index; Davis, 1983; Lucas-Molina et al., 2017) as well as evidence that women perform better than men at nonverbal reasoning and emotion detection (Briton & Hall, 1995; Buck et al., 2017; Hall et al., 2025). A growing body of research across constructs has argued that, rather than innate differences in "empathy," gendered socialization and stereotypes may be the cause of these gender differences (Klein & Hodges, 2001; Martingano et al., 2025). Evidence from inferential accuracy research supports this trend. When researchers do find evidence of gender differences, women outperform men on inferential accuracy tasks. However, differences between men and women are often *not* found (e.g., Hodges et al., 2011) and as with other areas of "empathy" research, these inconsistent findings have been attributed to differences

in motivation due to salience of a gender roles (Hodges et al., 2011; Ickes et al., 2000; Klein & Hodges, 2001).

**Motivation.** Continuing the discussion of motivation, in addition to the salience of gender roles influencing perceivers' motivation in inferential accuracy studies, Ickes and colleagues (1990) found evidence for increased inferential accuracy in heterosexual participants guessing the thoughts and feelings of attractive opposite-sex targets, reasoning that the attractiveness increased perceiver motivation to attend to the target. Evidence for other incentives increasing inferential accuracy via increased motivation is mixed – Klein and Hodges (2001) found evidence for money increasing accuracy using Ickes' standard stimulus paradigm. However, Berlamont and colleagues (2023) found that manipulating partner-serving motivation (the motivation to support one's partner's wellbeing) *decreased* heterosexual men's inferential accuracy for their romantic interactants in conflict interactions.

**Familiarity.** A sizeable portion of inferential accuracy research has been conducted specifically on romantic interactants (e.g., Hinnekens et al., 2021), and the relationship between perceiver-target familiarity and inferential accuracy is more complex than one may expect. Generally speaking, dyads who know each other well (e.g., romantic couples, friends) are more accurate than dyads of two strangers (Berlamont et al., 2024; Stinson & Ickes, 1992). However, some studies have found that inferential accuracy decreases over time in romantic relationships (Kilpatrick et al., 2002). Hinnekens and colleagues (2021) propose that this may be due to a transition in long-term relationships in which interactants move from using situation-specific cues to relying more on their general schema for the target (i.e., their romantic partner). For example, if one knows that one's spouse dislikes desserts that contain coconut, one may

inaccurately assume they dislike a savory dish because it has coconut milk, instead of attending to the situation-specific cues that they are enjoying the savory dish.

Additionally, Simpson and colleagues (1995) found evidence for a unique case of motivated *inaccuracy* in which romantic interactants may be motivated to be inaccurate in inferring their interactant's thoughts during a relationship-threatening situation (i.e., when their interactant was asked to make attractiveness ratings of attractive strangers). This effect was compounded for participants who were particularly insecure in their relationship.

**Accurate Stereotypes.** The relationship between inferential accuracy and use of stereotypes to make those inferences is contingent upon the accuracy of the stereotypes themselves. Lewis and colleagues (2012) examined perceivers' ability to infer the thoughts and feelings of new mothers and found that perceivers whose inferences contained accurate stereotypes about new mothers (e.g., that new mothers are tired or that new mothers love their babies) were more accurate than those who did not, suggesting that group-level knowledge was a useful tool for improving accuracy. Subsequent research by Hodges and Kezer (2021) supported this conclusion in a sample of American perceivers who inferred the thoughts of male Middle Eastern targets. Contrary to the researchers' hypothesis that negative stereotypes about this group of targets would impair inferential accuracy, results revealed that more stereotypical inferences once again actually predicted greater inferential accuracy, an effect attributed to the fact that stereotypes collected from the perceiver population were both less negative and more accurate than anticipated. Lewis and colleagues (2012) found evidence that stereotypical inferences were most related to greater accuracy when targets' thoughts were themselves more stereotypical, supporting the trend that stereotypes facilitate inferential accuracy primarily when they contain

valid generalizations about group members' actual experiences and perspectives (although interestingly, this effect did not replicate in Hodges and Kezer's work).

**Verbal Information.** As noted at the beginning, of the predictors that have been investigated in inferential accuracy research, a perceiver's utilization of verbal information to make accurate thought-feeling inferences has garnered the most robust support (Gesn & Ickes, 1999; Hall & Schmid Mast, 2007; Hodges & Kezer, 2021; Kraus, 2017[4]). Prior evidence suggests that accuracy is negatively affected if videos of the target are shown without verbal information (Kraus, 2017) or with unintelligible verbal information (Gesn & Ickes, 1999). Gesn and Ickes (1999), Hall and Schmid Mast (2007) and Kraus (2017) also examined perceiver accuracy for stimuli shown without visual information (i.e., participants only listened to the audio of a stimulus) and no research team found evidence that removing visual information from a stimulus negatively impacted accuracy. In fact, Kraus (2017) reported preliminary evidence that accuracy improved slightly for audio-only stimuli, suggesting that visual information (e.g., "body language") may be a distractor that could hinder accuracy. As Kraus's operationalization of accuracy used Likert scales for discrete emotions, this evidence suggests that the importance of verbal information may not be limited to the Ickes-type open-ended questions.

Recent work by Hodges and Kezer (2021) has offered the clearest insight into the relationship between verbal information and accuracy. In Hodges and Kezer's (2021) study, when individual video clips of interactions were coded for perceivers' use of the target's spoken

[4] In lieu of the more common Ickes paradigm, Kraus (2017) used perceiver-target agreement on Likert scales for 23 discrete emotions (e.g., happy, irritation) as an operationalization of inference accuracy.

words in the perceivers' inferences[5], perceivers whose inferences most resembled the target's verbal responses during the interaction were higher in inferential accuracy. This effect was amplified, not surprisingly, when targets reported thoughts that also resembled what the targets had said out loud during the interactions (i.e., where targets' thoughts and feelings were transparent; see Lewis et al., 2012), suggesting that accuracy is highest when targets verbalize their thoughts and feelings and perceivers use that verbal information to make thought-feeling inferences.

**Nonverbal Cues.** As the current investigation focuses on verbal cues used to infer a target's thoughts and feelings, the role of nonverbal cues as clues to targets' thoughts and feelings and the rationale for not including them merits discussion. Much of the research on nonverbal cues to a target's thoughts and feelings has focused on emotion detection: is this person scared or happily surprised? Are they crying tears of joy or tears of sorrow? In a series of studies that examined extreme emotions using the arguably more ecologically valid approach of naturalistic (as opposed to posed) stimuli, participants failed to reliably distinguish between photos of tennis players who had just won or lost high stakes matches (Aviezer et al., 2017). In video stimuli that included paralinguistic cues (i.e., vocal noises without intelligible words), participants were unable to differentiate families reuniting with deployed soldiers from people reacting to terror attacks or home invaders (Atias et al., 2019; Wenzler et al., 2016). Across an array of emotions, participants made cross-valence errors regardless of whether stimuli were still photos, silent videos, or videos with non-language vocalizations (Anikin et al., 2018).

[5] Scored on a 4-point scale ranging from 0 (the perceiver's inference did not match the target's verbal responses during the interaction at all) to 3 (the perceiver's inference very closely matched the target's verbal responses during the interaction).

The key for accurate emotion categorization seems to be context. In a study by Chen and Whitney (2019), removing all cues to a target's emotional state by blacking out the target's face and body did not impact accuracy for emotion detection as long as the contextual cues remained (e.g., the video showed the blurred target receiving a speeding ticket). In their review, Atias and colleagues (2021) propose that non-verbal emotional cues are not intrinsically meaningful but instead create a feedback loop with semantic context such that a target's speech reveals the context for their emotions and emotional cues elucidate the semantic content (e.g., by providing cues to detect sarcasm; see Bryant & Fox Tree, 2005; Matsui et al., 2016).

One area where the continued reliance on non-verbal cues may affect thought-feeling inferences is when perceivers have reason to distrust verbal information, such as when the target may be lying. In one prior inferential accuracy study by DesJardins and Hodges (2015), perceivers inferred the thoughts of targets who were assigned to be truthful or to be less truthful in a mock job interview. Perceivers were more accurate at inferring a target's thoughts and feelings when perceivers were naïve to the possibility of the target lying (but this was true only when the targets were actually being honest). "Informed" perceivers seemed to be on "high alert," and thus possibly less likely to trust what targets were telling them.

The allure of reading body language (i.e., an emotions-based approach to deception detection) in lieu of a more prosaic focus on verbal information has substantial negative societal repercussions. For example, police who focus on "body language" have a significantly higher deceit-bias (i.e., false positive rate – they think people they are questioning are lying when they are not) and are less accurate in detecting deception than those who rely on verbal information (Vrij, 2008). Further, in a meta-analysis by Hartwig and Bond (2011), many cues that are frequently endorsed as "giveaways" of deception in the cultural context of the United States of

America show poor utility as cues for actually detecting deception: averting eye contact and fidgeting one's leg are substantially weaker cues to dishonesty than verbal cues such as refusing to discuss a topic. Across multiple meta-analyses, evidence suggests that perceivers are not looking at the wrong body language cues but instead that good body language cues are rare and are rarely obvious – instead, language is a much better key to the puzzle of what is happening in other people's heads (Hartwig & Bond, 2011, 2014).

## Models of Interpersonal Accuracy

The evidence reviewed above illustrates a clear opportunity for investigation: verbal information is a robust predictor of inferential accuracy, and its role may have been underappreciated because prior research has focused on identifying accurate perceivers rather than on understanding the information environment in which inferences are made. Because both Ickes paradigms use video recordings of real conversations, this investigation has a unique opportunity to examine the full landscape of verbal cues available to perceivers: every utterance, by every speaker, across the entire conversation. To frame this shift in focus, I turn to two theoretical models of interpersonal judgment. Prior research, in predominantly seeking to identify predictors of accuracy (be they associated with the perceiver, the target, or the dyad), has fallen under the umbrella of Funder's realistic accuracy model (RAM; Funder, 1995). The current study returns to an older model, Brunswik's (1956) lens model. In the section that follows, I first review the RAM before contrasting it with the lens model.

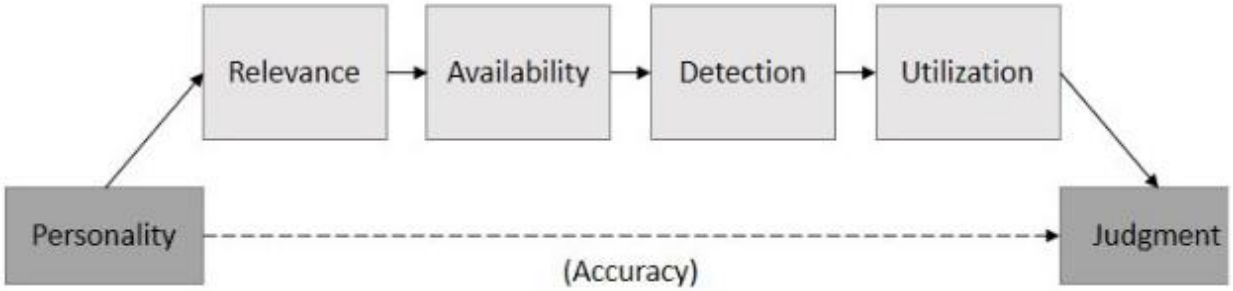
### *Funder's RAM*

Set in the tradition of personality perception, Funder's RAM (Figure 1) conceptualizes interpersonal accuracy in judging a latent construct in a target as a four-stage process in which first a person emits cues relevant to an internal state or trait. These cues are available to a

perceiver who detects them and utilizes the information to make an inference. For example, a perceiver may be tasked with inferring the level of extraversion in a target. For the perceiver to make an accurate inference, the target must emit cues relevant to their level of extraversion (e.g., engaging in conversation with a stranger) and these cues must be available to (i.e., seen by) the perceiver. The perceiver must then detect these cues (i.e., notice that the target is talking to a stranger) and utilize the information to make an accurate inference. An inaccurate utilization may result in the perceiver concluding that the target is only talking to the stranger to flirt with them.

**Figure 1:**

*Funder's (1995) Realistic Accuracy Model*




***The Lens Model***

In contrast to RAM is the lens model, which was developed by Egon Brunswik in the 1950s as a probabilistic model to describe judgment and decision-making (Brunswik, 1955, 1956). As described in Karelaia and Hogarth's (2008) meta-analysis, the judgment of a participant, $S\,(Y_s)$, results from a combination[6] of weights applied to a sample space of *j* proximal ecological cues where each cue $X_j$ is weighted $\beta_{s,j}$ as to its informativeness to the decision at hand (plus error, $\epsilon_s$) as shown in Equation 1[7]:

[6] In the traditional mathematical approach to the lens model (e.g., Tucker, 1964), the cues are combined linearly.

[7] Per APA 7 guidelines, equations in this dissertation will be labeled with sequential numbers shown in parentheses on the right-hand side of the page.

$$Y_s = \sum_{j=1}^{k} \beta_{s,j} X_j + \epsilon_s \quad (1)$$

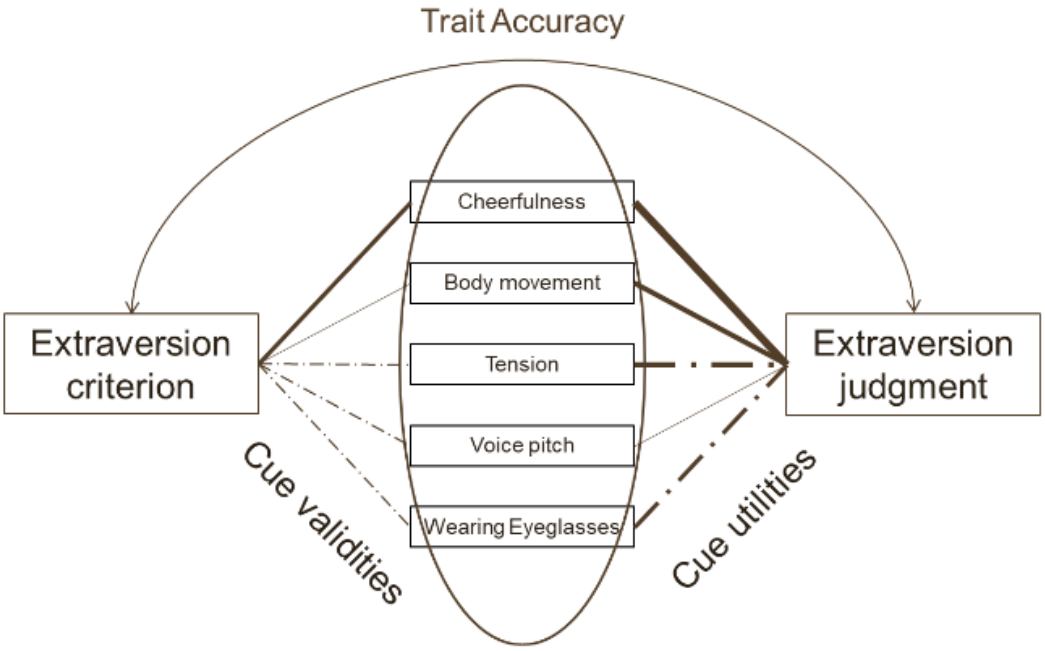
The ecologically valid inference, $Y_e$, (i.e., the correct inference) is modeled in Equation 2 (below) following the same formula in which each cue $X_j$ is weighted $\beta_e$ based on its validity as representative of the underlying construct.

$$Y_e = \sum_{j=1}^{k} \beta_{e,j} X_j + \epsilon_e \quad (2)$$

The degree to which perceiver weights $\beta_{s,j}$ correlate to the ecologically valid weights $\beta_{e,j}$ then constitutes the level of accuracy. Almost all studies utilizing the mathematical approach to the lens model have a numeric outcome such as a prospective employee's job performance (Dougherty et al., 1986) or numeric estimates of personality traits (Osterholz et al., 2019). Figure 2 shows an example of the lens model from Breil and colleagues (2021, p. 11) of a perceiver inferring a target's level of extraversion. In their model, the target's level of extraversion, labelled *extraversion criterion*, is connected to each accuracy cue by lines whose thickness represent $\beta_{e,j}$ (i.e., how representative a cue is of the target's extraversion). The weights $\beta_{s,j}$ (i.e., how much each cue was weighted by the perceiver) are shown in the thickness of the lines from each cue to the perceiver's extraversion judgment. Cheerfulness is the most representative cue to the target's extraversion and is the most utilized cue for the perceiver's judgment, representing proper calibration. However, body movement seems to have been improperly calibrated: although it was low in validity (i.e., not a strong representation of the target's extraversion, as indicated by the line representing $\beta_{e,body\ movement}$ being faint), the bold line connecting body movement to the perceiver's extraversion judgment suggests it was heavily utilized in perceivers' judgments.

**Figure 2:**

*Example Lens Model from Breil et al. (2021)*




The lens model has been used in social and behavioral research in several areas, including examining the development of rapport (Bernieri & Gillis, 2001; Bernieri et al., 1996); identifying what cues customers use when judging service or product quality (Bristow et al., 2015); and as a tool to model personality measurement through creative writing (Küfner et al., 2010). In one particularly salient example, Bhowmik and colleagues (2025) leveraged the lens model to identify what cues perceivers use to make academic-related judgments of students. Researchers first recorded pairs of high school students participating in science experiments. Afterward, perceivers watched the videos and attempted to infer several attributes about each student, including their academic self-concept, intelligence and motivation. Stimuli videos were then coded for a set of 17 cues, including dynamic cues such as the student's facial expressions

as they changed throughout the experiment and static cues such as the student's perceived gender or whether they wore glasses.

By analyzing a full sample space of available cues, Bhowmik and colleagues were able to identify patterns in how perceivers made judgments about students. Importantly, these patterns did not necessarily lead to accurate judgments but instead reflected trends in perceiver judgment (i.e., they were able to identify cue utilization irrespective of cue validity). For example, researchers found that students who wore eyeglasses were judged as significantly more intelligent and more intrinsically motivated than those who did not, whereas students who wore more distinctive clothes were judged to be more intrinsically motivated but not more intelligent – however, neither of these were *valid* cues. The 17 selected cues explained between 43% and 58% of the variability in perceiver judgments for each trait. Of particular interest to the current study, the application of the lens model by Bhowmik and colleagues (2025) allowed the researchers to not only examine how perceivers were making their judgments but the degree to which each cue increased the accuracy of each criterion judgment (e.g., a positive academic self-concept was associated with friendly behaviors in the video-recorded lab experiment).

### *Present Investigation*

The trio of studies in the current investigation is the first to apply the lens model to verbal cues in inferential accuracy. In doing so, I define a sample space of cues that includes every verbal utterance in each inferential accuracy stimulus video and examine several variables that are theoretically likely to predict which cues are most utilized by perceivers and which are most representative of what targets were actually thinking (i.e., valid). Studies 1 and 2 examine the right side of the lens model, cue utilization: what attributes of each verbal cue predict higher utilization by perceivers? Study 1 focuses on structural attributes (when in the conversation an

utterance occurs, who said it, and in what paradigm), while Study 2 focuses on discursive attributes, categorizing every utterance by its dialogue act type (e.g., a question, a suggestion, an informative statement). Study 3 turns to the left side of the lens model (cue validity), examining which cues are genuinely representative of target thoughts. Together, these studies offer an account of how verbal information functions as a lens through which perceivers infer, sometimes accurately, what another person is thinking and feeling.

Beyond these substantive questions, this dissertation contributes a novel methodological framework for studying verbal information in inferential accuracy. Prior investigations of what perceivers "use" when inferring a target's thoughts and feelings have established a robust relationship between verbal information and inferential accuracy using methods that did not attend to fine-grained information about which words were spoken when – for example, by removing all verbal information from a stimulus video (e.g., Kraus, 2017). To my knowledge, the present investigation is the first to apply modern natural language processing tools to inferential accuracy research and to create a pipeline for this process that future researchers can use.

As described in greater detail in Chapter II, I first used weakly-supervised, locally hosted algorithms to transcribe the inferential accuracy stimuli (Radford et al., 2022). I then used sentence-embedding models (Reimers & Gurevych, 2019) to represent everything said by targets and interactants, every perceiver inference, and every target thought as high-dimensional semantic vectors. The cosine between pairs of these vectors yielded a continuous, utterance-level similarity outcome indexing how closely each utterance aligned with what a perceiver inferred or a target reported thinking. Pairing this similarity outcome with Bayesian multilevel Gaussian process models (Bürkner, 2017; Rasmussen & Williams, 2008) then allowed me to examine

verbal information at a fine temporal resolution without imposing a priori assumptions about the shape of its influence across a conversation. Finally, I used individual participant data meta-analyses (Burke et al., 2017) to pool evidence across six previously collected studies while preserving the heterogeneity of their designs.

As far as I am aware, none of these tools have been leveraged previously to answer questions about how perceivers attend to and utilize verbal information. Earlier methods were not well-suited for asking, for instance, whether speech from particular speakers, at particular moments in a conversation was preferentially used by perceivers as they constructed their inferences. Such questions become tractable when transcripts broken down to the utterance level and modeled with the time-varying tools described above. Because all analytic components of this investigation are open source, I have also established a research pipeline that consolidates novel tools into a replicable template for future work seeking to quantify verbal cues in dynamic social interactions which could, with modest adaptation, be extended to adjacent interpersonal accuracy paradigms.

## II. STUDY 1: STRUCTURAL PREDICTORS OF PERCEIVER INFERENCES

In this study (Study 1), I analyze the temporal structure of social interactions to investigate whether and how the location of speech within a conversation, the speaker type (target vs. interactant), and the study paradigm (Ickes' standard stimulus vs. dyadic interaction) predicted the degree to which perceiver inferences matched transcriptions of speech uttered during the conversation. I do so by examining the structural predictors of perceiver inferences by modeling *semantic similarity*, defined as the cosine similarity between the embedding vectors of individual speech utterances from transcribed stimulus videos and the embedding vectors of perceivers' written inferences of targets' thoughts and feelings. In plain terms, higher values of speech-inference similarity indicate that what a perceiver wrote about their guesses of what the target was thinking more closely matched the words actually spoken in the conversation. Importantly, this operationalization assumes that similarity is reflective of attention to and utilization of the words spoken during the conversation but does not directly test it, as is discussed further in the General Discussion (Chapter V). Higher semantic similarity between an utterance and a perceiver's inference may reflect genuine attentional weighting, but it could also arise from shared conversational context, common cultural scripts, or other sources of convergence that do not require direct attention to specific utterances. Throughout the analyses that follow, I use the lens model terminology of "cue utilization" and "cue validity" as an organizing framework, while acknowledging that this interpretive mapping rests on the assumption that semantic similarity reflects something about how perceivers construct inferences.

Study 1 tests four hypotheses relating to the expected predictors of semantic similarity based only on these structural variables (cue utilization). Questions related to discursive content (i.e., dialogue acts made during the conversation) are examined in Study 2.

**Hypothesis 1:** Semantic similarity between words in the conversation and the target's reported thoughts and feelings will increase as conversation approaches the point at which the perceiver's inference is made.

**Hypothesis 2:** Perceiver inferences will show higher semantic similarity with target speech than with interactant speech, consistent with the theoretical claim that perceivers attend more to target speech than interactant speech.

**Hypothesis 3:** As speech occurs closer to the point of inference, interactant speech will show a weaker (but still positive) increase in semantic similarity than target speech as proximity to the inference is likely to increase similarity for the conversation as a whole.

**Hypothesis 4:** Interactant speech in the dyadic interaction paradigm will be more semantically similar to perceiver inferences than interactant speech in the standard stimulus paradigm, reflecting perceivers attending to their own speech as interactants.

This chapter also introduces the two-stage analytical method that will be used in Studies 2 and 3. In all studies, Stage 1 first fits individual Bayesian multilevel models with a Gaussian process function (to model the nonlinear trajectory of the conversations) on the raw data from each study[8]. In Stage 2, the estimates from these individual models are pooled in a series of Bayesian meta-analyses. Given the novelty of this Bayesian approach to time-series data in inferential accuracy research, Study 1's method section includes a much more comprehensive

[8] Within the text, "study" (with a lowercase "s") refers to the dataset from a previously collected inferential accuracy study. "Study" with a capital "S" refer to the three studies unique to this dissertation.

overview of the model equations, priors, and rationale than would be typical, in order to orient readers to these analytical methods. Studies 2 and 3 build upon this groundwork and thus are more abbreviated.

## Method

### *Included Studies*

I used pre-existing raw data from six inferential accuracy studies in this investigation. All data were collected as part of inferential accuracy research that has not previously been transcribed or used in this type of analysis. As several studies share authors and several of them are unpublished, I gave descriptive names to each dataset to improve ease of reading. Three studies were named after a particularly salient aspect of the targets' identity: In the *New Mothers* study, Lewis and colleagues (2012) used new mothers asked about their experiences being new mothers as targets[9] and in the *Middle Eastern Men* study, Hodges and Kezer (2021) asked Middle Eastern male students in higher education questions about their home country and experiences in the United States. Both studies used the standard stimulus paradigm to examine the role of stereotypes and stereotypicality between thoughts and inferences as sources of accuracy. In the *STEM* study, Hodges and colleagues (2022; unpublished dataset) recorded undergraduate students discussing their plans to enroll in graduate school with graduate students in science, technology, engineering and math (STEM) fields. The graduate students in this study acted as targets and the undergraduates acted as perceivers.

The other three studies recruited from the undergraduate population of the University of Oregon and were named for a unique aspect of the social interactions: in the *Round Robin (In*

[9] Stimulus videos for the New Mothers study were collected as part of an earlier study by Hodges and colleagues (2010), but the perceiver inferences used here were collected by Lewis and colleagues (2012).

*Person)* study, Lewis (2014; unpublished doctoral dissertation) had previously unacquainted dyads of undergraduate students participate face-to-face in a round-robin style set of introductory conversations during which they answered questions designed to elicit self-disclosure. In the *Round Robin (Video Mediated)* study, Kezer (doctoral dissertation, in prep) used the same round robin, get-to-know-you approach with undergraduate students, but the students interacted with each other online. Finally, in the *Negotiation* study, Howington (2016; unpublished doctoral dissertation) had undergraduates engage in negotiations for small prizes. Descriptions of each study are presented in Table 1.

**Table 1:**

*Included Studies*

| Short Name | Study | Summary |
|---|---|---|
| Middle Eastern Men | Hodges and Kezer (2021) | Middle Eastern male students studying in higher education settings in the U.S. were interviewed about their experiences in the United States and their home countries. |
| Negotiation | Howington (2016; unpublished doctoral dissertation) | Undergraduate students engaged in conversation about how they would divide up a set of small prizes. |
| New Mothers | Lewis et al. (2012) | New mothers whose first child was 2-4 months old were interviewed about their experiences as new mothers. |
| Round Robin (In Person) | Lewis (2014; unpublished doctoral dissertation) | Previously unacquainted undergraduate students answered questions designed to elicit self-disclosure in a round-robin set of face-to-face dyadic conversations. |
| Round Robin (Video Mediated) | Kezer (doctoral dissertation, in prep) | Previously unacquainted undergraduate students answered questions designed to elicit self-disclosure in a round-robin set of dyadic conversations held online. |
| STEM | Hodges et al. (2022; unpublished dataset) | Undergraduate students engaged in informal conversations with graduate students in STEM fields about the undergraduates' preparation and readiness for STEM graduate school. |

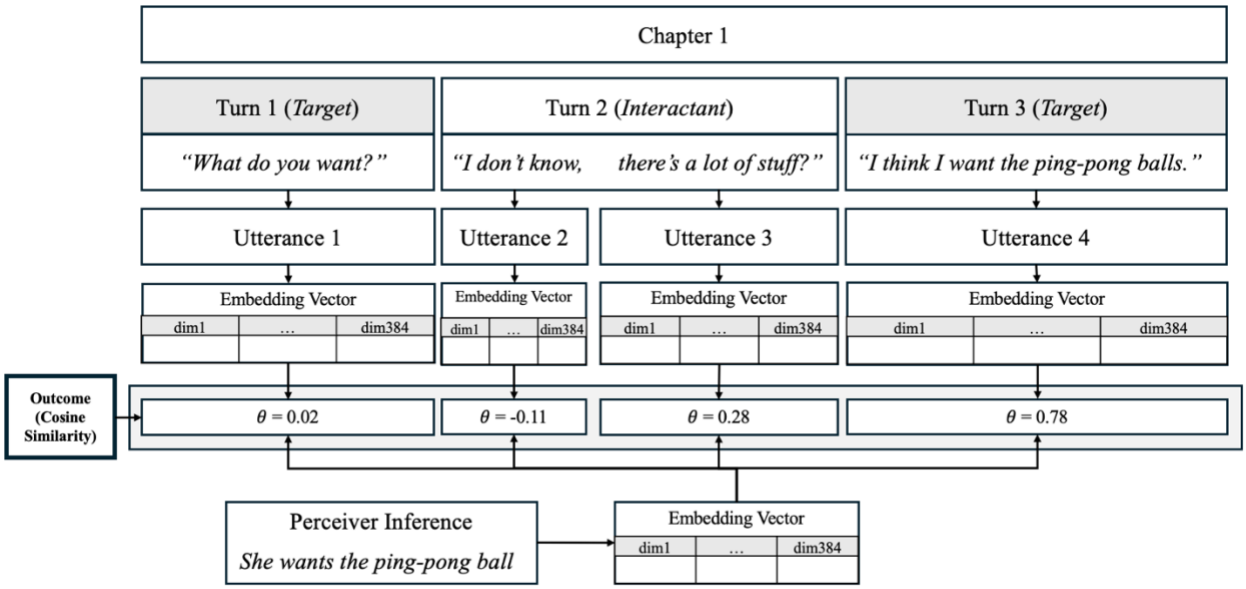
### *Data Preparation*

I initially transcribed each social interaction stimulus video using an open-source, weakly-supervised speech recognition model created by OpenAI (Whisper; Radford et al., 2022). These transcripts were then reviewed by research assistants who tagged speakers (i.e., targets and

interactants) and corrected the transcriptions for accuracy, resulting in 74.18 hours of transcribed video data. Each transcript was then broken into three nested levels: chapters, turns, and utterances (see Figure 3). The highest level, chapters, constituted the speech between points when the video was paused and the target reported their thoughts and feelings. Most studies had between four and eight chapters per stimulus target, though studies with target-specified inference schedules had a much broader range of inferences per video.

**Figure 3:**

*Example of Semantic Similarity*



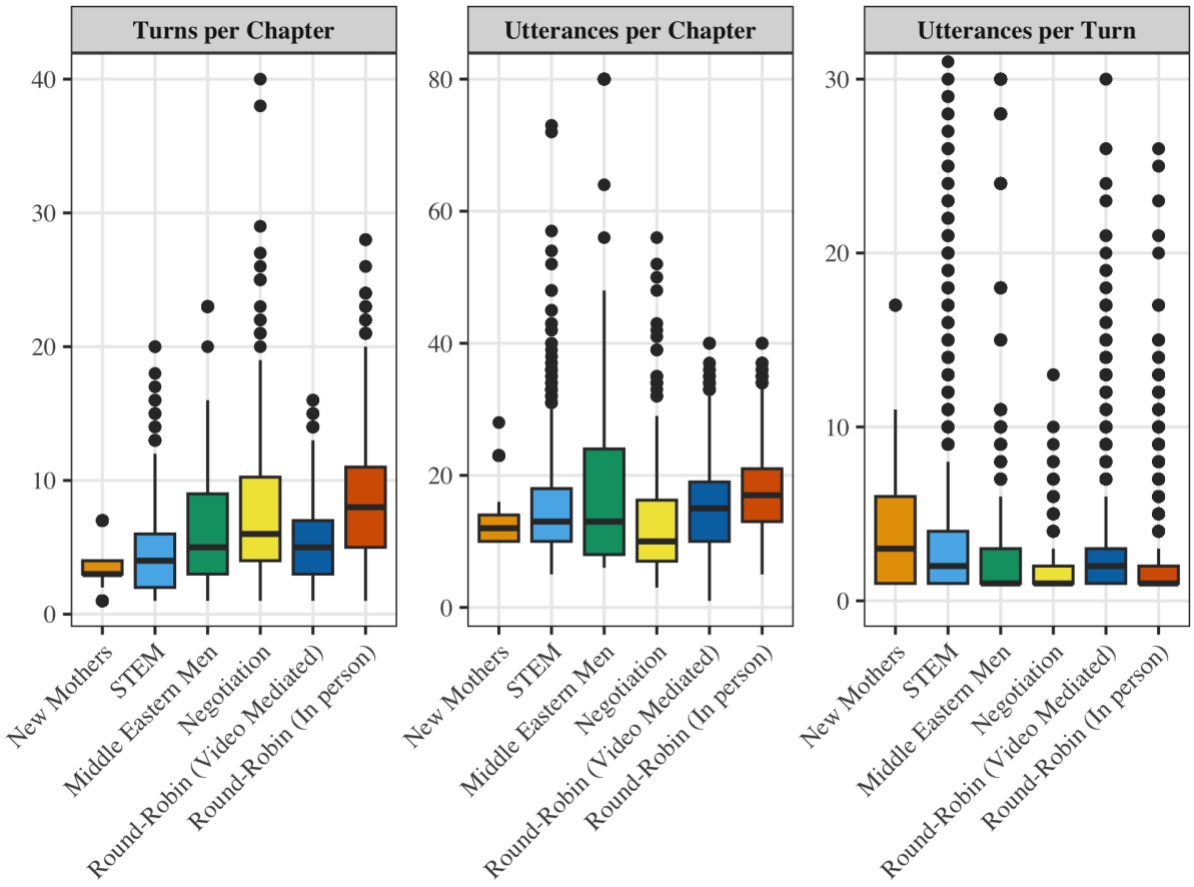

Chapters were then broken into turns, defined as an individual speaker's complete speech between another speaker's speech. If a stimulus video was viewed as a script, a turn would be equivalent to an actor's line – a turn may be a full monologue or a brief utterance. Although speech may intermittently overlap, the analyses largely utilized the speech of a single speaker and identified the start and end of a turn based on the timestamp of the video, not in reference to when another person began speaking.

The smallest units were called utterances[10]. Although there are variations in the specific delineation between discrete utterances across various research traditions, the current investigation operationalizes utterances as complete dialogue acts that could constitute meaningful turns if viewed in isolation (Stiles, 1992, 2017). To separate turns into discrete utterances, I hand-coded a subset of 5,000 utterances from a mixture of all studies which were subsequently used to train a branch of the SaT-6l model (Frohmann et al., 2024) that split turns into utterances based on the learned data. This resulted in 96.78% accuracy on a holdout sample of 500 utterances comparing researcher-coded and model-derived utterance boundaries. Descriptive statistics on chapter length and the number of utterances and turns per chapter are shown in Figure 4 and described in detail in Table 2.

[10] Utterances, discursive acts, and illocutionary acts are all, for the purposes of the current investigation, synonymous.

**Figure 4:**

*Frequencies of Utterances and Turns Across Studies*




**Temporal Indexing.** To examine how semantic similarity between perceiver inferences and transcribed speech changes across a chapter, each utterance was assigned a normalized time index. Specifically, each utterance's beginning timestamp within the chapter was divided by the total duration of that chapter, yielding a time value ranging from 0 (the start of the chapter, immediately after the prior inference point or the beginning of the video) to 1 (the moment at which the video was paused and the perceiver reported their inference). I opted for a proportional time scale over raw timestamps or utterance indices because it standardized chapters of differing durations onto a common metric, allowing the Gaussian process to estimate temporal dynamics

comparably across studies with different chapter lengths. In practical terms, a time index of 0.75 means that 75% of the chapter has elapsed; if a perceiver's inference aligns most strongly with an utterance at $t = 0.90$, that utterance occurred near the very end of the chapter, just before the video was paused. Because the time index is based on the timestamp of each utterance rather than its ordinal position, speech that is temporally close (e.g., even if it is a lengthy turn) receives similar time values that allow variability to be shared along the Gaussian process even if it spans multiple utterance indices, preserving real time structure of the conversational data.

**Table 2:**

*Descriptive Statistics for Conversational Structure Across Studies*

| | | | N | | | Chapter Length | Turns per Chapter | | Utterances per Chapter | | Utterances per Turn | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Study** | **Paradigm** | **Inference Schedule** | **Perceivers** | **Targets** | **Inferences** | **M (SD)** | **M (SD)** | **Range** | **M (SD)** | **Range** | **M (SD)** | **Range** |
| Middle Eastern Men | Standard Stimulus | Variable | 323 | 9 | 3,768 | 2:04 (1:43) | 5.83 (4.73) | [1, 23] | 18.13 (14.92) | [6, 80] | 3.11 (4.37) | [1, 30] |
| Negotiation | Dyadic Interaction | Variable | 41 | 41 | 224 | 2:06 (1:40) | 8.13 (6.35) | [1, 40] | 13.98 (10.08) | [3, 56] | 1.72 (1.22) | [1, 13] |
| New Mothers | Standard Stimulus | Variable | 90 | 14 | 1,198 | 1:04 (0:23) | 3.09 (1.34) | [1, 7] | 13.02 (3.33) | [10, 28] | 4.21 (3.57) | [1, 17] |
| Round Robin (In Person) | Dyadic Interaction | 45s | 537 | 537 | 2,555 | 2:08 (1:03) | 8.83 (4.54) | [1, 28] | 17.47 (6.35) | [5, 40] | 1.98 (1.71) | [1, 32] |
| Round Robin (Video Mediated) | Dyadic Interaction | 45s | 195 | 195 | 2,523 | 2:05 (1:04) | 5.47 (2.92) | [1, 16] | 15.18 (6.23) | [1, 40] | 2.77 (2.56) | [1, 30] |
| STEM | Dyadic Interaction | 45s | 160 | 160 | 1,638 | 5:30 (2:58) | 4.22 (2.87) | [1, 20] | 14.59 (7.17) | [5, 73] | 3.45 (4.28) | [1, 73] |

*Note. Mean Chapter Length and standard deviations are presented as minutes:seconds.*
*Inference Schedule refers to whether the points at which targets reported their thoughts and feelings were self-selected by the targets themselves (Variable) or predetermined by the researchers to occur every 45 seconds (45s).*

**Semantic Similarity.** In order to quantitatively analyze the qualitative language data, I had to create a numeric representation of semantic similarity. This was accomplished by first transforming transcribed speech into numeric data using natural language processing (NLP) models. NLP models map text to a multi-dimensional vector space, creating a vector of numeric data for each utterance and inference that could be compared to quantify the amount of semantic similarity between two pieces of text. These vectors, called embeddings[11], were generated using the all-MiniLM-L6-v2 model within Sentence-Transformers (Reimers & Gurevych, 2019), which transformed the text into embedding vectors of 384 numeric data points for each discrete utterance. The all-MiniLM-L6-v2 model was adapted from the pretrained MiniLM-L6-H384-uncased model (Wang et al., 2020) to be a sentence and short paragraph encoder trained on one billion sentence-pairs scraped from various sources, including Reddit, Stack Exchange, and Wikihow.

Semantic similarity was then operationalized using the cosine of the angle between the NLP embedding vector of a perceiver's inference and the embedding vector of a unique utterance from one of the transcribed videos. Figure 3 (above) shows a simplified, simulated example of calculating cosine similarity between speech turns. In this example, each turn of transcribed speech is shown above a vector of variables (i.e., embeddings) determined via an NLP algorithm. The similarity between each utterance and the perceiver inference is shown as a cosine similarity score below the turn embeddings. In the simulated example, the perceiver inference[12], "She wants the ping pong ball" is most similar to the utterance, "I think I want the

[11] Embedding vectors consist of empirically derived variables that describe unique features of the text but lack theoretical meaning and are only used in their full form.
[12] When later testing the exploratory hypotheses in Study 3 examining which aspects of speech correlate with target thoughts, the process was repeated substituting the perceiver inference embeddings with the target-reported thoughts.

ping pong balls" with a semantic similarity (i.e., cosine similarity) value of 0.78. As can be seen in Table 3 and Figure 5, cosine similarity between perceiver inferences and target/interactant speech spanned a wide range. Semantic similarity was generally positively skewed: although some utterances had a similarity of over 0.90 (nearly identical to the perceiver's inference), most utterances by both targets and interactants had a cosine similarity of between 0.10 and 0.20 with perceiver inferences.

Although infrequent, some cosine similarity values were negative, indicating utterance–inference pairs whose embedding vectors pointed in opposing directions in the semantic space. However, because sentence-transformer embeddings tend to occupy a narrow positive cone in the vector space (e.g., Zhou et al., 2022), a more useful understanding of these negative values is that they reflect a near-total absence of shared semantic content rather than a meaningful semantic opposition or antonymy (Steck et al., 2024).

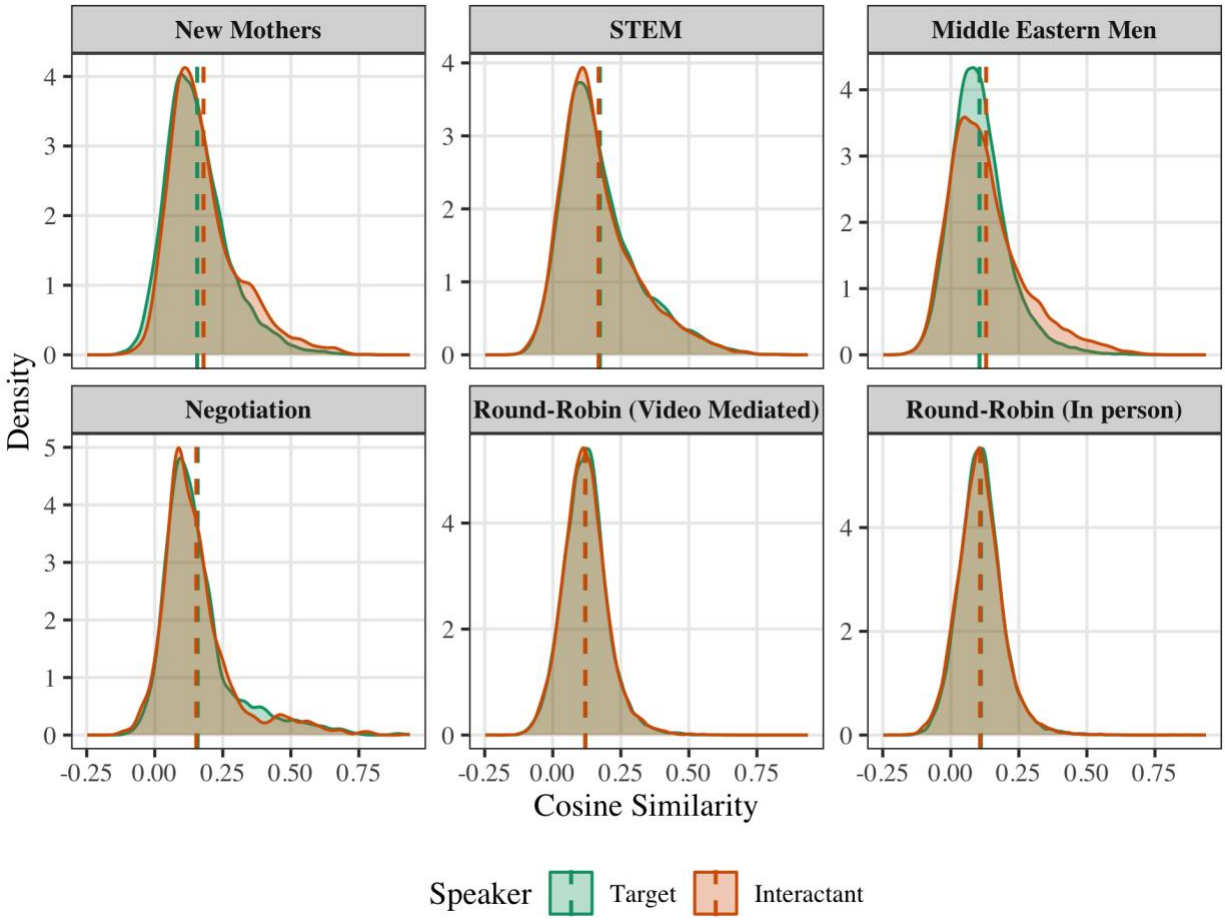
**Table 3:**

*Mean, Standard Deviation, and Range of Cosine Similarity Across Studies*

| | ***Target*** | | ***Interactant*** | |
|---|---|---|---|---|
| ***Study*** | ***M (SD)*** | ***Range*** | ***M (SD)*** | ***Range*** |
| Middle Eastern Men | 0.105 (0.105) | [−0.248, 0.834] | 0.13 (0.138) | [−0.177, 0.757] |
| Negotiation | 0.158 (0.135) | [−0.1, 0.905] | 0.153 (0.136) | [−0.128, 0.932] |
| New Mothers | 0.156 (0.117) | [−0.146, 0.786] | 0.179 (0.13) | [−0.122, 0.779] |
| Round Robin (In Person) | 0.111 (0.081) | [−0.139, 0.875] | 0.108 (0.083) | [−0.218, 0.822] |
| Round Robin (Video Mediated) | 0.120 (0.082) | [−0.215, 0.654] | 0.119 (0.082) | [−0.168, 0.608] |
| STEM | 0.173 (0.141) | [−0.149, 0.937] | 0.168 (0.14) | [−0.148, 0.827] |

**Figure 5:**

*Distribution of Cosine Similarity Between Targets and Perceivers Across Studies*




***Analytic Overview: Two-Stage Individual Participant Data Meta-Analysis***

The analytic process used in all three studies in this dissertation follows that of an individual participant data (IPD) meta-analysis (Riley et al., 2010, 2021). In a conventional aggregate-data meta-analysis, researchers pool published summary statistics; in an IPD meta-analysis, the original data from each contributing study are analyzed directly, preserving participant-level detail. As there are only six studies in this investigation and the data are already nested in several levels, I implemented a two-stage variant of IPD meta-analysis rather than include study as a grouping variable (e.g., Burke et al., 2017; Debray et al., 2015). In Stage 1, I

modeled each of the six studies individually using a Bayesian multilevel Gaussian process model, yielding study-specific parameter estimates. In Stage 2, I pooled these estimates across studies using both a Bayesian random-effects meta-analysis and a synthesized trajectory analysis based off of the pointwise posterior distributions identified by the Gaussian process. Although the one-stage IPD analysis would be preferable, this two-stage approach accommodates the substantial design heterogeneity across the six studies (described in greater detail below) without requiring the complex cross-study random-effects specifications that a one-stage model would demand. The analyses still benefit from the IPD methods as all analyses as well as data cleaning and transformation methods are identical, minimizing unobserved variability due to researcher decision-making.

I chose the two-stage approach over a one-stage joint model for several reasons. First, the six studies differ substantially in design (standard stimulus vs. dyadic interaction), sample size (ranging from 3,294 to 68,300 utterances), number of targets, and chapter structure. A one-stage model pooling all studies simultaneously would require extremely complex random-effects specifications to accommodate these differences, with questionable identifiability for the Gaussian process components. Burke and colleagues (2017) showed that two approaches typically yield similar results except when studies are very small, there are very few studies, or when modeling strongly non-linear relationships. Although these conditions may seem characteristic of this dataset ($k = 6$, strongly non-linear within-study trajectories), the two-stage approach is appropriate here because the GP captures within-study non-linearity adequately, leaving Stage 2 to pool a manageable set of summary parameters. (for practical guidance on the implementation of IPD meta-analysis, see Debray et al., 2015).

### *Stage 1: Individual Bayesian Multilevel Gaussian Process Models*

As the analytical methods I used are novel for this area of research, Study 1 includes a much more thorough explanation of the modelling process than is typical; Studies 2 and 3 build off of this foundation and describe only the ways in which the models differ from those in Study 1. This explanation of the modelling process begins with a review of the core likelihood function, then discusses the nested structure of the multilevel model, and ends with a description of the prior selection process and post hoc sensitivity analyses. Each study is reported using the specifications and guidelines offered by the Bayesian Analysis Reporting Guidelines (BARG; Kruschke, 2021).

**Model Equation.** The likelihood distribution, linear model, and priors of the multilevel models used to test the hypotheses in Study 1 are described here. Given the complexity of the multilevel model, the full model is presented in Equation 3, and each part of the equation is described individually below (for more information on Bayesian estimation, see Etz et al., 2018).

$$
\begin{aligned}
Y_{ijkl} &\sim \text{Normal}\left(\mu_{ijkl}, \sigma\right) \\
\mu_{ijkl} &= 0 + \beta_{s_{ijkl}} + f_{s_{ijkl}}\left(t_{ijkl}\right) + u_{l,s_{ijkl}} + v_{jk,s_{ijkl}} \\
f_s &\sim \mathcal{GP}\left(0, k_{\text{Matérn}_{3/2}}\right) \text{ for } s \in \{\text{target,interactant}\} \\
u_{l,s} &\sim \text{Normal}\left(0, \tau_{\text{perceiver},s}\right) \text{ for } s \in \{\text{target,interactant}\} \\
v_{jk,s} &\sim \text{Normal}\left(0, \tau_{\text{target:chapter},s}\right) \text{ for } s \in \{\text{target,interactant}\} \\
\beta_s &\sim \text{Normal}(0.14, 0.15) \\
\alpha &\sim \text{Exponential}(20) \\
\ell &\sim \text{Inverse-Gamma}(3, 0.5) \\
\tau_{\text{perceiver},s} &\sim \text{Exponential}(10) \\
\tau_{\text{target:chapter},s} &\sim \text{Exponential}(10) \\
\sigma &\sim \text{Exponential}(10)
\end{aligned}
\tag{3}
$$

**Likelihood Function.** The multilevel model used in all studies begins with a likelihood function around the outcome, cosine similarity. In the first line of the model in Equation 4, cosine similarity is represented as $Y_{ijkl}$, and can be read as "the cosine similarity for utterance $i$ in chapter $j$ of target $k$ as observed by perceiver $l$ is distributed along a normal distribution around $\mu_{ijkl}$ and $\sigma$."

$$
Y_{ijkl} \sim \text{Normal}\left(\mu_{ijkl}, \sigma\right) \tag{4}
$$

**Linear Multilevel Model.** The next line of the equation is shown in Equation 5 and defines the expected cosine similarity between an individual utterance and a perceiver inference

($\mu_{ijkl}$) as a linear combination of parameters. The $\beta$ coefficients for the intercepts of each speaker type are determined using cell-means parameterization[13], in which the implied intercept for the entire linear model is set to 0 and a fixed effect for each speaker type (target or interactant) is estimated using the group's mean cosine similarity. This can be considered similar to "dummy codes" that allow the model to estimate a unique intercept for targets and interactants, increasing the interpretability of the findings. As such, two fixed effects $\beta_{\text{target}}$ and $\beta_{\text{interactant}}$ were estimated (Equation 5). Intercept $\beta_{\text{target}}$ multiplied by the indicator function[14] $\mathbb{1}[S_{ijkl} = \text{target}]$ is multiplied by 1 when speaker $S_{ijkl}$ = target is true and by 0 when speaker $S_{ijkl}$ is the interactant (and thus the indicator function is then false). The corresponding fixed effect for interactant ($\beta_{\text{interactant}}$) is multiplied by the inverse indicator function: $\beta_{\text{interactant}}$ is multiplied by 1 (and thus included in the linear model) when speaker $S_{ijkl}$ is the interactant and 0 when $S_{ijkl}$ is the target.

$$\mu_{ijkl} = 0 + \beta_{\text{target}} \cdot \mathbb{1}[S_{ijkl} = \text{target}] + \beta_{\text{interactant}} \cdot \mathbb{1}[S_{ijkl} = \text{interactant}] + f_{S_{ijkl}}(t_{ijkl}) + u_{l,S_{ijkl}} + v_{jk,S_{ijkl}} \quad (5)$$

***Gaussian Process Temporal Dynamics***

Gaussian processes allow researchers to estimate nonlinear trends in temporal dynamics without committing to a specific parametric form (Rasmussen & Williams, 2008). This is done by estimating time-varying deviations from the mean using shared variance across data points.

[13] This is contrasted with the more common "reference" coding approach $\mu = \beta_0 + \beta_1 \cdot \mathbb{1}[S = \text{interactant}]$ in which $\beta_0$ generally refers to the mean of the reference group (i.e., targets) and $\beta_1$ represents the difference between targets and perceivers.

[14] Indicator functions are a type of numeric Boolean function that is set to 1 when the internal condition is true and 0 when the internal condition is false.

For example, a researcher studying how bone density changes over the lifespan may employ a sequential design in which they recruit multiple cohorts of different ages whose bone density is tracked for three years. Although a participant at age 28 may have very similar bone density at age 31, they are less likely to have a similar level of bone density at age 80. Further, the rate of bone density gain or loss may change with age — bone density may remain relatively stable for many years but change dramatically (e.g., when a participant experiences menopause). The Gaussian process in this context would allow researchers to estimate how similar bone density at time point $t$ is expected to be at time point $t'$ and the prior expected rate of change and amplitude of deviations from the expected mean are determined by the covariance kernel (described below). Returning from the analogy to the current investigation, the Gaussian process here models how semantic similarity between the conversation stream and perceiver inferences changes across the chapter: Are the first few utterances of a conversation just as aligned with the perceiver's eventual inference as the final ones, or does similarity build gradually?

Research using conversational data in particular benefits from Gaussian process modeling tools. There is considerable disagreement between researchers on how to split text into 'utterances,' and even trained researchers using the same system can vary considerably in determining what counts as a distinct utterance (e.g., Biron et al., 2021). However, in leveraging models with Gaussian processes, proximal utterances are allowed to share variability (i.e., trends in the Gaussian process curves may extend over several utterances). As one could imagine, utterances near one another are likely to be more similar than those further apart because proximal utterances often are discussing the same topic or have a logical connection as part of the conversational flow. The amount that proximal utterances do (or do not) share variability is thus examined in this study as an indicator of conversational heterogeneity. Thus, researcher-

level differences in operationalizing and delineating utterances are somewhat mitigated as, for example, two utterances that I coded as discrete but that another researcher may argue should be collapsed into a single utterance can share variability due to their temporal proximity.

In Equation 6, the Gaussian process *f* is defined uniquely for each chapter's target and interactant speech. It has two arguments, the mean function (0) and the covariance kernel ($k$). In this model equation, the mean function is set to 0, corresponding to an a priori belief that, at any given time point, the Gaussian process should estimate a mean of 0. However, because the mean cosine similarity for targets and interactants is already estimated in the fixed effects of β_target and β_interactant, the mean function being set to 0 does not imply that the function expects a mean cosine similarity of 0, but instead that it does not expect deviation from the already-estimated mean. The other argument, the covariance kernel ($k$) is described in the next section.

$$f \sim \mathcal{GP}(0, k) \tag{6}$$

**Matérn 3/2 Covariance kernel.** The second argument in the Gaussian process is the covariance kernel (k). In the current study, models were specified using a Matérn 3/2[15] covariance kernel (see Equation 7; Stein, 1999) as it allows for functions to be once differentiable[16], thus allowing a moderate degree of roughness in estimating similarity (i.e., it does not over-smooth the data as would happen in twice differentiated kernels) while not bouncing too drastically from moment-to-moment. The equation defining $k_{\text{Matérn}_{3/2}}(t, t')$

[15] Sensitivity analyses (discussed below) tested the influence of kernel specification on posterior distributions.

[16] As in calculus, being once differentiable refers to the smoothness of the slope: the Matérn 3/2 kernel is smooth enough that it is possible to calculate its slope (first derivative) at any point, but not necessarily smooth enough to reliably calculate how quickly that slope itself is changing (second derivative).

specifies that the expected similarity between time point $t$ and a second time point $t'$ is defined by two hyperparameters: amplitude (α) and length-scale ($\ell$).

$$f_s(t) \approx \sum_{j=1}^{m} \sqrt{S_{\text{Matérn}_{3/2}}\left(\sqrt{\lambda_j}\right)}\, \phi_j(t)\, \beta_j, \quad \beta_j \sim \text{Normal}(0,1) \tag{7}$$

Amplitude α corresponds to the marginal standard deviation of the fluctuations from the mean function; for example, an amplitude of 0.10 would suggest that approximately 68% of the Gaussian process function values are expected to fall within 0.10 of the mean function (and thus the mean cosine similarity for that target/interactant); the larger the amplitude, the more the Gaussian process function is expected to "wiggle." The length-scale (ℓ) specifies how quickly the function can change and thus how much of the function is correlated; a useful heuristic for the length-scale in this study (which has a time scale ranging from 0 to 1.0) is that the value of ℓ roughly corresponds to what proportion of the chapter would have to progress before the conversation becomes uncorrelated. In other words, a ℓ of 0.10 would suggest that utterances within about 10% of the chapter are strongly correlated, or that approximately 10 unique (and thus unrelated) conversational sections occurred within that one chapter.

To summarize: amplitude captures the *magnitude* of temporal fluctuation (a large amplitude means speech-inference similarity varies a lot across the chapter) and length-scale captures the *smoothness* of that fluctuation (a small length-scale means the conversation shifts topics rapidly; a large length-scale means the conversation maintains a consistent thread). These two hyperparameters, along with the fixed effects (mean speech-inference semantic similarity), are the quantities estimated in the Stage 1 models and subsequently pooled across studies in Stage 2.

As it pertains to the language data in these studies, the general shape of the Matérn 3/2 kernel allows for asides or rapid deviations from the general conversational flow to influence the data without the Gaussian process jumping every turn. A target briefly stating that they felt ill before returning to casual pleasantries would be highlighted in these data if the perceiver emphasized that specific utterance to make their inference, whereas similar moments may be smoothed over by other kernels (for a discussion of other Gaussian process kernels, see Wilson & Adams, 2013). In other words, this allows for single-turn utterances, such as an interactant asking a specific question, to influence the data without bouncing around at every turn.

**Hilbert Space Gaussian Process (HSGP).** The large samples and nested nature of this study necessitated the use of Hilbert Space Gaussian Process (HSGP) approximation with k = 25 basis functions (Riutort-Mayol et al., 2023). Computational complexity in traditional Bayesian Gaussian process modeling results in $O(N^3)$ transformations (i.e., order of N cubed) because the N x N covariance matrix must be inverted as part of every step in the Markov Chain Monte Carlo (MCMC) sampling process. The HSGP approximation allows for basis functions (using Laplace eigenfunctions) to approximate this process via a set of functions derived from the spectral representation of the covariance kernel (Solin & Särkkä, 2020). In other words, the infinite-dimension Gaussian process is broken into k = 25 smooth basis functions that jointly approximate the full Gaussian process function. Although a relatively novel application in the specific context of this investigation, the use of this approximation in other structurally similar studies can result in less than a 1% error[17] rate compared to the true Gaussian process estimation,

[17] 1% error is a more-than-acceptable tradeoff for the 439 trillion operations that would occur while modeling the utterances in Lewis (2014) alone, a process that would take over six months to run on a modern computer.

though approximation accuracy is sensitive to the number of basis functions and boundary factor relative to the function's length-scale (Riutort-Mayol et al., 2023).

**Random Effects.** Two random effects are specified in Equation 8 and allow each perceiver to systematically weight target and interactant speech independently[18]. Specifically, $u_{l,\text{target}}$ captures how much each perceiver deviates from the overall mean for target speech (defined as a normal distribution around 0). Similarly to the mean function in the Gaussian process, the mean of normal distribution set at 0 means it does not differ from the overall average specified by $\beta_{\text{target}}$. The normal distribution has a standard deviation $\tau_{\text{perceiver, target}}$ which quantifies how much the perceivers differ from each other.

$$\begin{aligned} u_{l,\text{target}} &\sim \text{Normal}\left(0, \tau_{\text{perceiver, target}}\right) \\ u_{l,\text{interactant}} &\sim \text{Normal}\left(0, \tau_{\text{perceiver, interactant}}\right) \end{aligned} \tag{8}$$

A second set of random effects (Equation 9) is specified that allows each chapter *j* within target/interactant *k* to have its own deviation. In this case, the random effects allow the specific cosine similarity from one chapter to deviate from the average cosine similarity for all target or interactant speech across all chapters as specified in $\beta_{\text{target}}$ and $\beta_{\text{interactant}}$. Parallel to the previous random effects, as the mean cosine similarity of target or interactant speech is estimated by the fixed effect $\beta_{\text{target}}$ or $\beta_{\text{interactant}}$, centering the normal distribution on 0 keeps the modal estimate for target or interactant similarity at 0 (i.e., no different from the fixed effect estimate). The component $\tau_{\text{target:chapter, target}}$ also captures the heterogeneity of chapters within a target. A large $\tau_{\text{target:chapter, target}}$ suggests that cosine similarity differed sizably between chapters, whereas a small $\tau_{\text{target:chapter, target}}$ implies that the chapters had little variability.

[18] An alternative model with correlated random effects is included in the sensitivity analysis.

$$v_{jk,\text{target}} \sim \text{Normal}\left(0, \tau_{\text{target:chapter, target}}\right)$$
$$v_{jk,\text{interactant}} \sim \text{Normal}\left(0, \tau_{\text{target:chapter, interactant}}\right) \quad (9)$$

Finally, all remaining variability is defined by $\varepsilon_{ijkl}$ (i.e., the residual variance) is estimated as a normal distribution around 0 with a standard deviation of $\sigma$.

***Prior Specification***

Priors in Bayesian analysis are numeric representations of the researcher's beliefs prior to seeing any data. As this is an entirely novel approach to data in this context, weakly informative priors (Table 4) were specified for this (and all subsequent) Bayesian models.

**Table 4:**

*Primary Priors for Bayesian Multilevel Models*

| Parameter | Description | Prior |
|---|---|---|
| $\beta_s$ | Mean cosine similarity for speaker type $s$ | $\text{Normal}(0.14, 0.15)$ |
| $f_s(t)$ | GP function for temporal dynamics of speaker $s$ | $\mathcal{GP}\left(0, k_{\text{Matérn}_{3/2}}\right)$ |
| $\alpha$ | GP marginal standard deviation (amplitude) | $\text{Exponential}(20)$ |
| $\ell$ | GP length-scale | $\text{Inverse-Gamma}(3, 0.5)$ |
| $u_{l,s}$ | Perceiver random effect for speaker type $s$ | $\text{Normal}\left(0, \tau_{\text{perceiver},s}\right)$ |
| $v_{jk,s}$ | Target:Chapter random effect for speaker type $s$ | $\text{Normal}\left(0, \tau_{\text{target:chapter},s}\right)$ |
| $\tau_.$ | Random effect standard deviations | $\text{Exponential}(10)$ |
| $\sigma$ | Residual standard deviation | $\text{Exponential}(10)$ |

The fixed effect (i.e., the expected cosine-similarity level) for targets and interactants was specified as a normal distribution centered on the grand mean of cosine similarity in the data (0.14) with a standard deviation of 0.15. More intuitively, the fixed effect priors expect approximately 95% of cosine similarity scores to fall between −0.16 and 0.44. An exponential(10) prior was chosen for all random effects. This places greater probability on small values while allowing for larger estimates if supported by the data. In other words, little variability from the grand mean is expected unless well supported by the data.

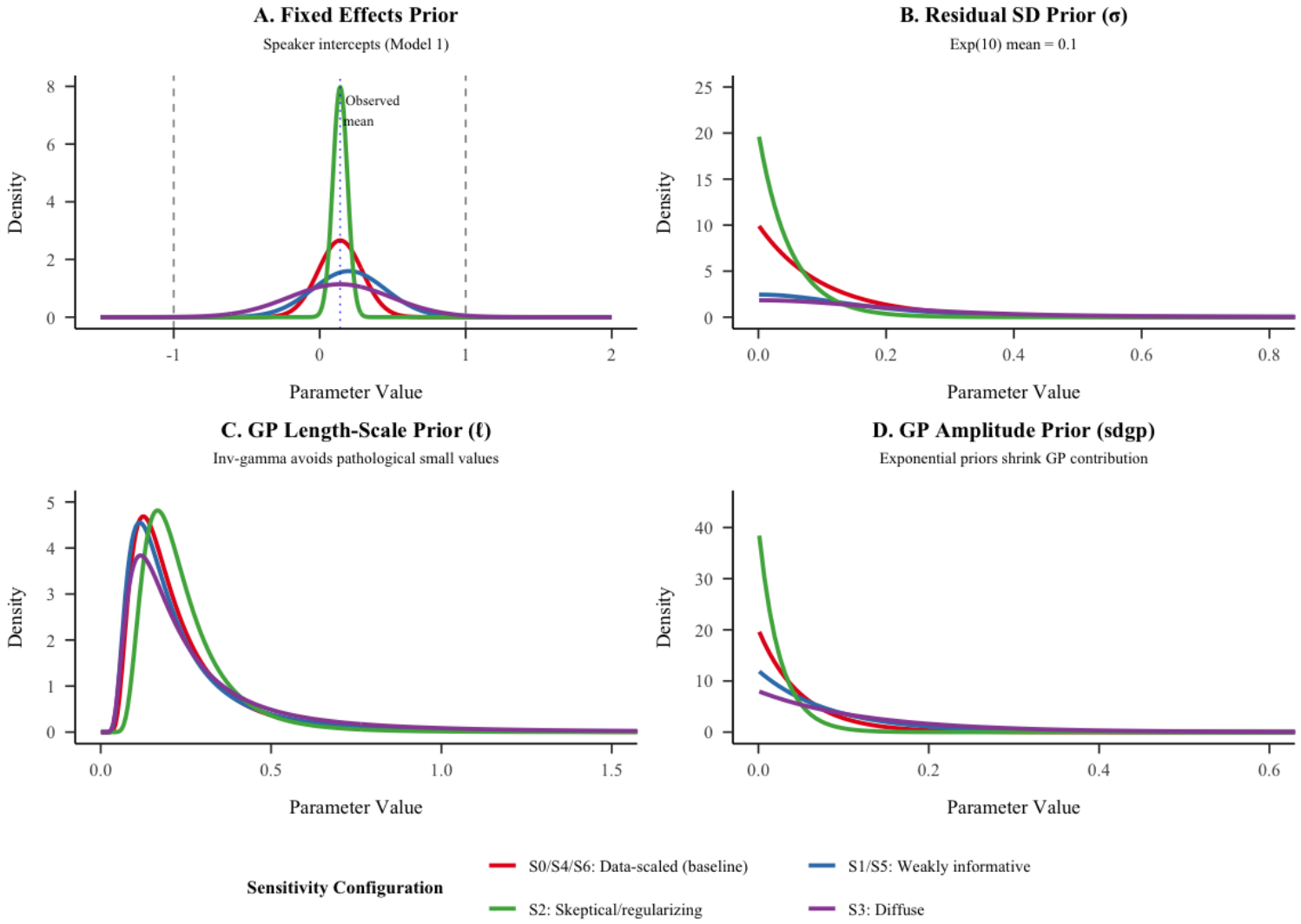
For the Gaussian processes, the distribution of amplitude priors ($\alpha$) was specified as exponential(20), encouraging smooth trajectories in cosine similarity over time by minimizing temporal noise. The length-scale priors used an inverse-gamma(3, 0.5) distributions, which enforces positive values (as length-scales cannot be negative) and places minimal mass near zero (Gelman et al., 2017).

***Sensitivity Analyses***

A natural concern when specifying priors is that, given insufficient data, they may unduly influence the findings. To mitigate this, in addition to the priors for the primary model (S0), six alternative models (S1-S6) with different priors were run in conjunction with the main model and their outcomes were compared to identify whether changing the priors influenced the conclusions. The visual distributions of these priors on the expected estimates for each parameter are shown in Figure 6.

**Figure 6:**

*Visualizations of Prior Distributions Used in Sensitivity Analyses*



The six priors test three different types of influence. The first three (S1-S3) test the effect of priors that encourage or limit regularization; S4 and S5 test the influence of a different GP kernel; and S6 adds correlation matrices to the primary model. All priors are shown in Table 5. Of the regularizing/diffuse priors, S1 is the most similar to S0, although slightly less "informative" (i.e., it does not expect semantic similarity to be as close to the mean); S2 regularizes more towards the grand mean of semantic similarity, and S3 is the most diffuse, allowing the data to spread to either limit of cosine similarity [−1, 1]. S4 and S5 test a different Gaussian process kernel, the squared exponential kernel, which is much smoother than the Matérn 3/2 kernel. All sets of priors S0-S5 estimate no correlation structure (i.e., random effects are independent). In S6, LKJ(2/4/2) places an LKJ prior on the correlation matrix for each level of the model, with more regularization LKJ(4) on the random effects of target (as the level with the fewest observations) and less regularization on the random effect of chapter and perceiver.

***Table 5:***

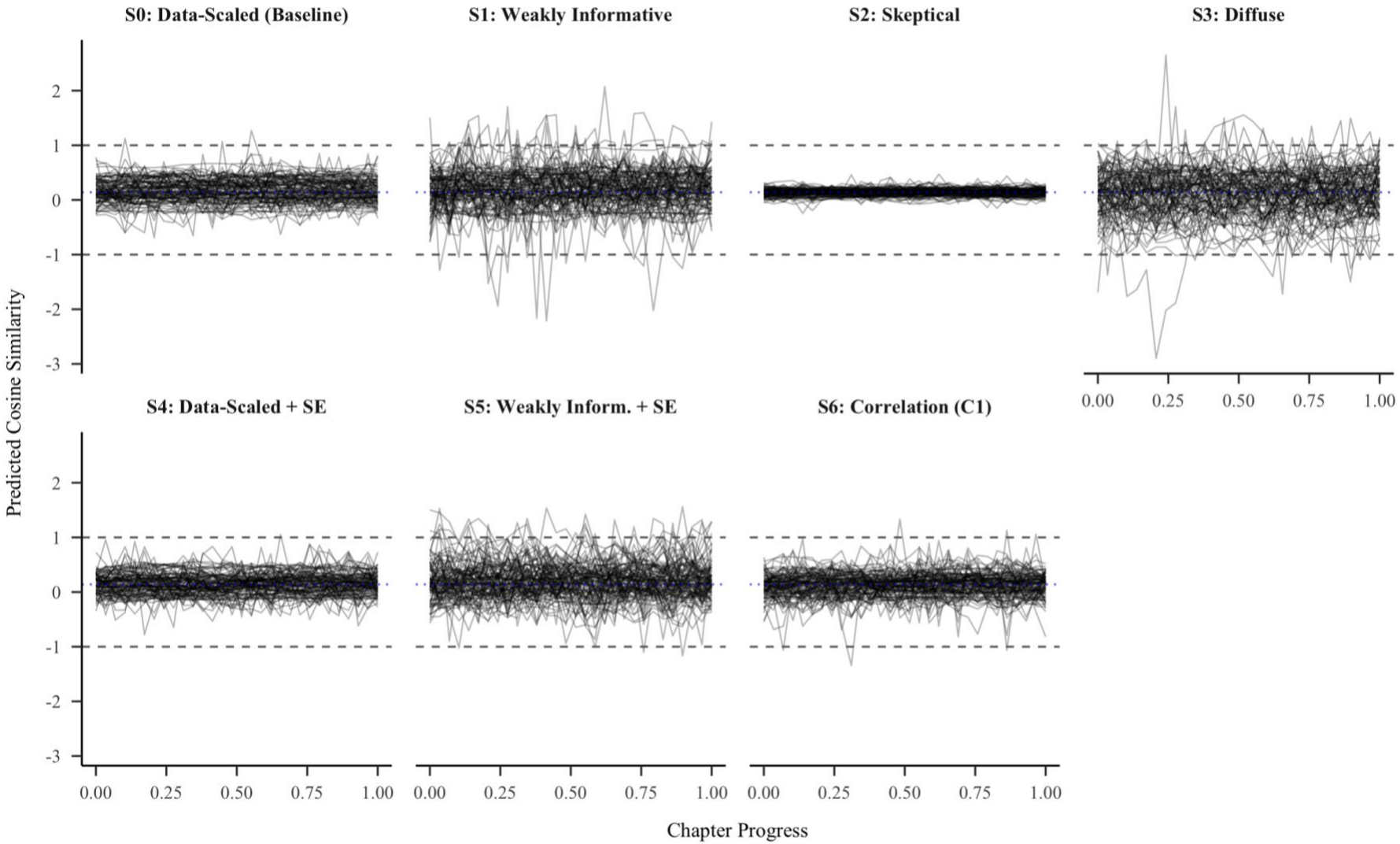
*Alternative Priors for Sensitivity Analyses*

| Priors | Fixed Effects | Resid. SD | Random Effects SD | GP Kernel | Length-scale (GP) | Amplitude (GP) | Cor. |
|---|---|---|---|---|---|---|---|
| *S0 (Primary)* | N(0.14, 0.15) | Exp(10) | Exp(10) | Matérn 3/2 | InvGamma(3, 0.5) | Exp(20) | None |
| *S1: Weakly Informative* | N(0.20, 0.25) | $t_3$(0, 0.15) | $t_3$(0, 0.15) | Matérn 3/2 | InvGamma(2.5, 0.4) | Exp(12) | None |
| *S2: Skeptical* | N(0.14, 0.05) | Exp(20) | Exp(20) | Matérn 3/2 | InvGamma(5, 1.0) | Exp(40) | None |
| *S3: Diffuse* | N(0.14, 0.35) | $t_3$(0, 0.20) | $t_3$(0, 0.20) | Matérn 3/2 | InvGamma(2, 0.35) | Exp(8) | None |
| *S4: Baseline + Squared Exponential* | N(0.14, 0.15) | Exp(10) | Exp(10) | Squared Exp. | InvGamma(3, 0.5) | Exp(20) | None |
| *S5: Weakly Inf. + Squared Exponential* | N(0.20, 0.25) | $t_3$(0, 0.15) | $t_3$(0, 0.15) | Squared Exp. | InvGamma(2.5, 0.4) | Exp(12) | None |
| *S6: Correlation* | N(0.14, 0.15) | Exp(10) | Exp(10) | Matérn 3/2 | InvGamma(3, 0.5) | Exp(20) | LKJ(2/4/2) |

Although the visualizations of each parameter's priors shown in Figure 6 may be informative to some readers, and each individual prior has a meaningful interpretation, the interaction of multiple prior distributions may have unintended repercussions on the actual data implied by the priors. As such, I passed a simulated dataset to each set of priors and drew 7,500 posterior estimates from those prior-only models. Subsamples of these estimates are shown in Figure 7, which may be useful in interpreting the range of regularization and diffusion, with S2 expecting a far more limited range of semantic similarity across the chapter than S3. Given that cosine similarity has a clear upper and lower bound [−1, 1], the proportion of draws from each set of priors that fell outside the expected range of the data was used to ensure the priors were appropriately specified. For the most diffuse priors, S1, S3 and S5, between 97% and 98% of draws fell within the appropriate range, and (after rounding to three decimal places), 100% of draws fell within the expected range for all other sets of priors.

**Figure 7:**

*Posterior Data Predicted by Alternative Priors in Sensitivity Analyses*



### *Model Fit*

As part of the Bayesian modeling process, two types of convergence statistics were used to determine the quality of the posterior sampling. These include $\hat{R}$[19], the potential scale reduction factor, which generally is viewed to reflect successful convergence when under 1.01 (for a discussion of model convergence criteria, see Vehtari et al., 2021). The second metric for determining convergence is large effective sample sizes (ESS). ESS is broken into two sub-statistics: bulk ESS measures the sampling sizes from the center of the distribution and can be viewed as an indication of sufficient representation of the modal ranges of data within the models. Tail ESS acts as a complement to bulk ESS and represents sampling at the edges of the data (i.e., where data are thinnest).

Although perspectives vary in terms of what constitutes sufficiently high ESS (e.g., Vehtari et al., 2021), I used 400 as an absolute minimum ESS for interpreting any models. However, as a general rule, models that didn't have a bulk or tail ESS of 1,000 also had Rhat values greater than 1.01, suggesting more foundational estimation issues. In most cases, this was due to insufficient observations, which primarily occurred when speech was categorized into dialogue acts in Chapter III. As the variables for cosine similarity and time (in percentage of the chapter) were already standardized, no adjustments to the data or priors were made to improve model convergence. In less technical language, if the modeling process is viewed as driving through a neighborhood to map the posterior data, the ESS and Rhat statistics reflect the gas mileage of the car – although not directly related to where the car ends up (i.e., the posterior estimations) they are informative in telling whether the route being taken is or is not efficiently covering the territory.

[19] Also called "Rhat"

**Markov Chains Monte Carlo.** Bayesian models use Markov Chain Monte Carlo (MCMC) to explore the multidimensional space informed by the data and priors to estimate a posterior distribution that is the basis for statistical insights (see McElreath, 2020). Stan (Carpenter et al., 2017), a probabilistic coding language that was used in R with the brms package (Bürkner, 2017), uses Hamiltonian Monte Carlo methods with a NUTS (No-U-Turn-Sampler) variant to improve efficiency in exploring the multidimensional posterior space. Given the many demands on these models (large samples, complex Gaussian processes, multilevel structure), all models were tested conservatively to ensure sufficient exploration of the sample space. This resulted in 4 chains running 8,000 iterations with 4,000 warmups (resulting in 16,000 samples) and an adaptive delta of 0.99[20].

***Stage 2A: Univariate Meta-Analysis of Individual Parameters***

In Stage 2A, I pooled the study-specific parameter estimates from Stage 1 using Bayesian random-effects meta-analysis. I meta-analyzed each parameter class (fixed effects, GP amplitude, GP length-scale, random effect standard deviations, and residual standard deviation) separately. The meta-analytic model treats each study's estimate as a draw from a normal distribution centered on the true population parameter, with known within-study standard error. Between-study heterogeneity is captured by $\tau$, the between-study standard deviation, which received a half-Cauchy(0, 0.5) prior (Röver, 2020; Röver et al., 2021).

Although I had access to the original data and could have used a one-stage approach, the complicated nesting structure, Gaussian processes, and small $k = 6$ meant an alternative approach was necessary to generate pooled estimates. I thus followed recommendations from Lunn and

[20] The adaptive delta determines the size of step taken by the MCMC chain, with 0.80 being the standard and 0.99 being substantially smaller, thus providing a more detailed exploration of the sample space.

colleagues (2013), who provided evidence that leveraging the standard errors estimated by original single-study models as known sampling variances allowed for robust meta-analytic pooling in contexts with similar limitations. GP amplitude and length-scale estimates were log-transformed before meta-analysis to satisfy the normality assumption of the random-effects model. The within-study standard errors on the log scale are obtained via the delta method (Oehlert, 1992; Ver Hoef, 2012), using the posterior standard deviation of the original-scale parameter.

**Likelihood Function**. For each study $i \in \{1, ..., 6\}$, speaker $s \in \{\text{target, interactant}\}$, and spline coefficient $k \in \{1, ..., K\}$, the observed coefficient from Stage 1 is modeled in Equation 10, where $y_{isk}$ is the posterior mean of the spline coefficient extracted from the individual brms model(s) and $\sigma_{isk}$ is the corresponding posterior standard deviation, treated as a known sampling error.

$$y_{isk} \mid \theta_{isk}, \sigma_{isk} \sim \text{Normal}(\theta_{isk}, \sigma^2_{isk}) \tag{10}$$

**Linear Model.** The true (latent) coefficient $\theta_{isk}$ (Equation 11) is decomposed into a population mean and study-specific deviation, where $\mu_{sk}$ represents the population mean for speaker $s$ and coefficient $k$, and $u_{i,sk}$ captures the deviation of study $i$ from this population mean. The population means $\mu_{sk}$ are the primary quantities of interest, as they characterize the 'typical' trajectory shape across studies.

$$\theta_{isk} = \mu_{sk} + u_{i,sk} \tag{11}$$

**Random Effects Structure.** The benefit of the joint Bayesian approach is the specification of study-level random effects as a single multivariate normal distribution that spans both speaker types. For each study *i*, the $2^{21}K$-dimensional vector of random effects is distributed as shown in Equation 12 where Σ is a $2K \times 2K$ covariance matrix. Following the standard parameterization in brms, this covariance matrix is decomposed such that D is a diagonal matrix of standard deviations ($\tau_1, ..., \tau_{2k}$) and R is a correlation matrix. This decomposition facilitates the specification of separate priors on the scale of heterogeneity (the τ parameters) and the correlation structure (R).

$$u_i = (u_i,\text{target},1, ..., u_i,\text{target},K, u_i,\text{interactant},1, ..., u_i,\text{interactant},K)^T \sim MVN_{2k}(0, \Sigma) \quad (12)$$

The covariance matrix Σ captures three distinct sources of association: (a) correlations among spline coefficients within target speech, (b) correlations among spline coefficients within interactant speech, and (c) correlations between target and interactant coefficients within the same study. This third component—the cross-speaker covariance—is what enables proper uncertainty quantification for speaker contrasts. In a study like the Negotiation study with relatively high semantic similarity for target and interactant speech, failing to model this positive correlation may lead to inflated variance estimates for the target-interactant difference.

**Prior Specification.** Weakly informative priors were specified for all parameters, following the general guidance of Gelman and colleagues (2017).

[21] K-dimensions multiplied by two speaker types (target, interactant).

***Population Means.*** The population mean coefficients were assigned the same priors as the individual models based on the grand mean of cosine similarity in the data (Equation 13).

$$\mu_{sk} \sim \text{Normal}(0.14, 0.15) \quad (13)$$

***Heterogeneity Standard Deviations.*** The between-study standard deviations were assigned half-Cauchy priors (Equation 14). As in the individual models, the half-Cauchy distribution was chosen because it has a peak at zero (reflecting a prior expectation of modest heterogeneity) while maintaining heavy tails that allow for large values if supported by the data (Gelman, 2006; Polson & Scott, 2012). The scale parameter of 0.5 was chosen based on the expected magnitude of between-study variation in trajectory coefficients.

$$\tau_j \sim \text{Half-Cauchy}(0, 0.5) \text{ for } j = 1, ..., 2K \quad (14)$$

***Correlation Matrix.*** The correlation matrix R was assigned an LKJ prior: $R \sim \text{LKJ}(2)$. The LKJ distribution (Lewandowski et al., 2009) is the standard prior for correlation matrices in Bayesian hierarchical models. With $\eta = 2$, this prior slightly favors correlations closer to zero while allowing strong correlations if warranted by the data. This choice sought to balance the naivety towards the correlation structure while providing sufficient regularization to ensure a positive-definite covariance matrix.

***Stage 2B: Multivariate Synthesis of Pointwise Posterior Trajectories***

As described earlier, the complex nature of the initial models, paired with the limited number of studies ($k = 6$) prevented use of a single meta-analysis model with all parameters. Stage 2A used univariate meta-analysis models to pool the estimates for each parameter and in Stage 2B, I recreated the Gaussian process by creating distributions across the pointwise posterior trajectories.

**Dimension Reduction Via Basis Function Coefficients.** Before synthesizing the pointwise posterior trajectories, each study's posterior predictive draws were transformed into a set of basis function coefficients. I accomplished this by using each fit model to take 500 posterior draws at 50 evenly spaced time points spanning the duration of the estimated Gaussian process. In other words, 500 posterior draws were taken every 2% of the Gaussian process function, creating a set of distributed estimates for each time point for each function. This process was repeated twice: once using the Gaussian process estimated for targets in each study and once using the Gaussian process estimated for interactants in each study.

This was still too many time points for a meta-analysis to pool across studies: to directly pool the trajectories from 50 time points would necessitate estimating a 50 x 50 covariance matrix, which was unfeasible for a meta-analysis of k = 6 studies. Following evidence collected by Belias and colleagues (2022), projecting each unique study-by-speaker trajectory onto a lower-dimensional basis function space allowed me to reduce the dimensionality of the covariance matrix substantially from over a thousand covariance parameters to 20. This was accomplished using natural cubic splines to create basis function representations of the Gaussian process function. I provide a brief explanation here; for a more comprehensive introduction, see Hastie and colleagues (2009) and Harrell (2015).

A natural cubic spline is a piecewise polynomial function that I used to break the smooth Gaussian process function into a set of cubic equations. It has four key attributes: it is piecewise cubic (it occurs between any two adjacent knots[22] ξ_j and ξ_{j+1}); it is continuous on the interval [*a,b*] (meaning that it spans the length of each chapter from [0,1]); the first and second derivatives are continuous at each interior knot; and it has linearity constraints that force data

[22] A knot is a breakpoint in the function

beyond the observed range to be linear[23]. A basis function then builds the d-degree polynomials that make up the intra-knot trajectories. An intuitive representation is that a basis function is like a linear model but the beta weights for each variable control how much of a specific curve to include in a shape and the combinations of these weighted curves is used to estimate the shape of the nonlinear process (see Belias et al., 2022)

The complexity of these meta-analytic models was then determined by the degrees of freedom of the natural cubic spline, calculated as the number of knots (change points) + 2. The more knots, the more inflection points occur in the natural cubic spline and thus more change points that can be modeled. Following the guidance provided by Harrell (2015), a range of knots were tested in a set of sensitivity analyses. These include three knots (three inflection points set at the 10$^{th}$, 50$^{th}$, and 90$^{th}$ percentiles), five knots (10$^{th}$, 20$^{th}$, 50$^{th}$, 80$^{th}$, 90$^{th}$) to 7 knots (at the 2.5$^{th}$, 18.33$^{rd}$, 34.17$^{th}$, 50$^{th}$, 65.83$^{rd}$, 81.67$^{th}$, and 97.5$^{th}$ percentiles). The number of knots had no strong influence on the material conclusions from Study 1 (all percentage overlaps > 95.72%). As such, the 3-knot natural cubic spline (with 5 degrees of freedom) is used here.

[23] The linearity constraint is used here because it minimizes erratic behavior at the edges of the prediction interval, improving estimation at the boundaries of the data (i.e., at the beginning and end of chapters).

Recalling that 500 posterior draws were taken at 50 time points from each study's Gaussian process, an ordinary least squares regression was fit predicting the trajectory values from the spline basis functions as shown in Equation 15.

$$\hat{y}_{isd}(t) = \sum_{k=1}^{K} \beta_{isdk} \cdot B_k(t) + \varepsilon \quad (15)$$

In the equation, $Bk(t)$ represents the $k^{th}$ natural cubic spline basis function evaluated at time $t$, and $\beta_{isd}$ is the 5-dimensional coefficient vector (five natural cubic spline basis coefficients). This approach resulted in coefficient vectors for all 500 posterior draws.

**Trajectory Reconstruction.** Pooled trajectories were reconstructed by multiplying the estimated population mean coefficients by the basis matrix evaluated at a fine grid of 100 time points as shown in Equation 16, where $B(t) = (1, B_1(t), ..., B_{k-1}(t))^T$ is the $K$-dimensional basis vector at time $t$. Crucially, this computation was performed at each of the 8,000 posterior draws, yielding a full posterior distribution for the trajectory at each time point.

$$\hat{f}_s(t) = B(t)^T \hat{\mu}_s \quad (16)$$

**Variance Ratio Diagnostic.** The spline projection step uses each study's posterior mean trajectory rather than the full posterior distribution. This approximation is justified when between-study heterogeneity substantially exceeds within-study posterior uncertainty, so that the added complexity of propagating the full posterior would negligibly affect the pooled estimates. I assessed this using a variance ratio diagnostic: for each spline coefficient, I computed the ratio of between-study variance (from the Stage 2A heterogeneity estimates) to within-study posterior variance (from the individual study posteriors) and used a ratio of 3 (or greater) to indicate that between-study variability dominates and the posterior-mean approximation is adequate.

**Hypothesis Testing Framework.** I used the synthesized trajectories to address the four hypotheses. I tested Hypothesis 1 (temporal increase) by computing the endpoint change from $t$ = 0 *t*o t = 1 and the median derivative across the trajectory, reporting the posterior probability of a positive value (i.e., a non-zero change in semantic similarity over time). For Hypothesis 2 (speaker differences), I compared target and interactant trajectories via pointwise contrasts (i.e., examining if their slopes differed). Hypothesis 3 (rate of change) examined whether the temporal derivative differs between speakers. Hypothesis 4 (paradigm moderation) re-fit the Stage 2B model with paradigm (dyadic interaction vs standard stimulus) as a study-level moderator. Throughout the results, "CI" refers to Bayesian credible intervals (the range within which the parameter falls with 95% posterior probability), not frequentist confidence intervals. For each hypothesis, I report the posterior probability that the parameter falls in the hypothesized direction, denoted Pr(Direction), computed as the proportion of posterior samples consistent with the predicted effect.

The evidence ratio converts this probability into an odds scale: Evidence Ratio = Pr(Direction) / (1 − Pr(Direction)), representing how many times more probable the hypothesized direction is relative to the alternative. Evidence ratios above 3 are interpreted as moderate evidence, above 10 as strong evidence, and above 30 as very strong evidence (Kass & Raftery, 1995). Because the prior probabilities for each directional hypothesis are set to .50 (reflecting no prior directional preference), the reported evidence ratios are numerically equivalent to Bayes factors under equal prior odds (Kass & Raftery, 1995; Wagenmakers, 2007). Because each hypothesis addresses a distinct theoretical question, no formal multiplicity correction was applied; however, exploratory analyses involving multiple contrasts should be interpreted with appropriate caution.

## Results

I applied the analytic pipeline described in the above Method section to six studies comprising 10,906 perceiver inferences drawn from a combined total of over 200,000 individual utterances. As a reminder, the outcome of interest was semantic similarity: the cosine similarity between the words spoken in a conversation and the words perceivers used to describe what the target was thinking. Higher values indicate more similarity between what was said and what perceivers inferred. Results are organized by analytic stage: Stage 1 reports individual study model fits and parameter estimates, Stage 2A reports pooled univariate meta-analytic estimates, and Stage 2B reports the synthesized posterior trajectories and hypothesis tests.

### *Stage 1: Individual Study Results*

**Convergence Diagnostics.** Table 6 shows the individual model fit statistics, with all models achieving $\hat{R} < 1.01$ and with minimum ESS (for both tail and bulk estimates) well over the preferred 1,000 ESS threshold, meeting model fit criteria.

**Table 6:**

*Study 1: Bayesian Multilevel Model Convergence Statistics*

| Study | N utterances | $\hat{R}$ (max) | ESS (min) |
|---|---|---|---|
| Middle Eastern Men | 68,300 | 1.00 | 1,938.28 |
| Negotiation | 3,294 | 1.00 | 6,727.80 |
| New Mothers | 15,603 | 1.00 | 6,199.15 |
| Round Robin (In Person) | 44,630 | 1.00 | 5,979.36 |
| Round Robin (Video Mediated) | 38,307 | 1.00 | 4,229.01 |
| STEM | 23,892 | 1.00 | 6,640.61 |

**Fixed Effects.** Fixed effect estimates (mean cosine similarity between utterances and perceiver inferences) are shown in Table 7. Across all six studies, both target and interactant speech showed positive mean cosine similarity with perceiver inferences, generally close to the grand mean of 0.14.

**Table 7:**

*Study 1: Individual Model Fixed Effect Estimates with 95% CIs*

| Study | Speaker | Estimate | SE | 95% CI | $\hat{R}$ | Bulk ESS | Tail ESS |
|---|---|---|---|---|---|---|---|
| Middle Eastern Men | Interactant | 0.08 | 0.09 | -0.1, 0.26 | 1.00 | 2,451.94 | 4,456.09 |
| | Target | 0.10 | 0.03 | 0.04, 0.16 | 1.00 | 1,938.28 | 3,753.13 |
| Negotiation | Interactant | 0.17 | 0.05 | 0.07, 0.27 | 1.00 | 7,081.35 | 6,751.21 |
| | Target | 0.17 | 0.04 | 0.09, 0.25 | 1.00 | 6,727.80 | 6,113.96 |
| New Mothers | Interactant | 0.22 | 0.07 | 0.07, 0.35 | 1.00 | 7,948.16 | 8,269.73 |
| | Target | 0.17 | 0.03 | 0.1, 0.24 | 1.00 | 6,542.54 | 6,677.53 |
| Round Robin (In Person) | Interactant | 0.10 | 0.02 | 0.04, 0.14 | 1.00 | 6,700.74 | 5,844.14 |
| | Target | 0.11 | 0.00 | 0.11, 0.12 | 1.00 | 5,979.36 | 3,929.95 |
| Round Robin (Video Mediated) | Interactant | 0.12 | 0.00 | 0.11, 0.13 | 1.00 | 5,838.25 | 3,354.37 |
| | Target | 0.12 | 0.00 | 0.12, 0.13 | 1.00 | 4,229.01 | 2,623.00 |
| STEM | Interactant | 0.17 | 0.02 | 0.13, 0.2 | 1.00 | 6,640.61 | 5,901.87 |
| | Target | 0.16 | 0.03 | 0.09, 0.22 | 1.00 | 7,164.93 | 5,947.55 |

Table 8 shows the estimated mean difference in semantic similarity between speakers (target vs. interactant). Across all studies, targets and interactants did not show robust within-study differences.

**Table 8:**

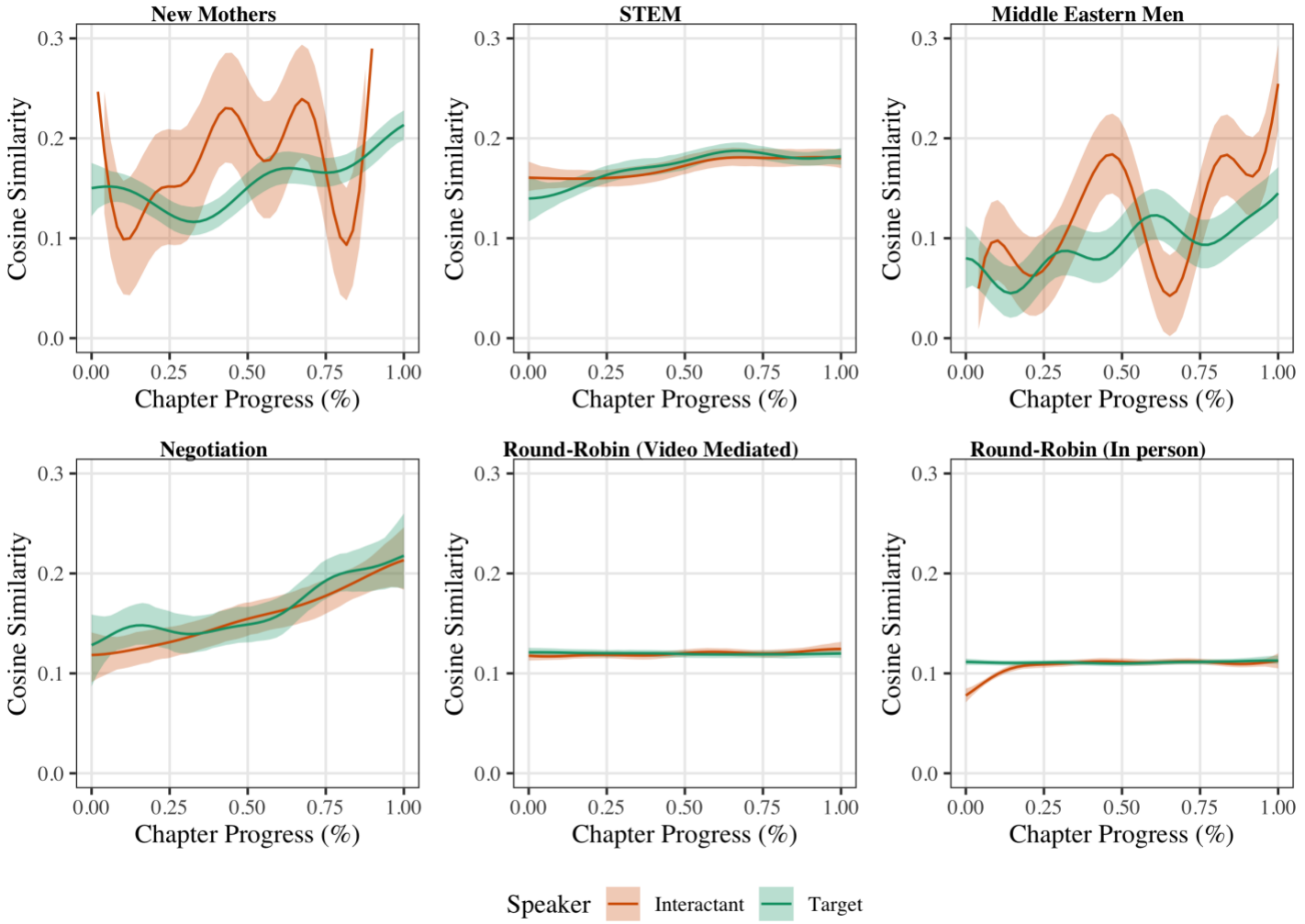
*Study 1: Speaker Contrasts in Semantic Similarity by Study with 95% CIs*

| Study | Mean difference (95% CI) | Probability (Target > Partner) |
|---|---|---|
| Middle Eastern Men | 0.02( -0.17, 0.21) | 0.58 |
| Negotiation | 0.00 (-0.12, 0.12) | 0.52 |
| New Mothers | -0.05 (-0.2, 0.11) | 0.25 |
| Round Robin (In Person) | 0.01 (-0.03, 0.07) | 0.80 |
| Round Robin (Video Mediated) | 0.00 (-0.01, 0.01) | 0.47 |
| STEM | -0.01 (-0.09, 0.06) | 0.42 |

However, visually examining the pointwise posterior trajectory estimates (Figure 8) suggests that the two speakers may be differentiated in ways not captured by the initial hypothesis of a main effect for speaker type. The Round Robin (In Person), Negotiation, and STEM studies did not show clear differences in semantic similarity between speakers, but they did all have varying upward trends, with semantic similarity seeming to increase over the course of the chapter. New Mothers and Middle Eastern Men, the two standard stimulus paradigm studies had clear differences between interactant and target semantic similarity for the last approximately 10% of the chapter, and the Round Robin (Video Mediated) study had a substantially lower semantic similarity for interactant speech in the first approximately 15% of the chapter. These observations are discussed more in context to the hypothesis tests below.

**Figure 8:**

*Individual Studies' Estimated Semantic Similarity by Speaker Over Time*



**Gaussian Process Hyperparameters.** GP amplitude and length-scale estimates are shown in Table 9. Amplitudes (the magnitude of temporal fluctuation in speech-inference similarity) varied substantially across studies: The two standard stimulus paradigm studies, New Mothers (interactant $\alpha = 0.26$) and Middle Eastern Men (interactant $\alpha = 0.39$), showed the largest amplitudes, indicating that speech-inference similarity for interactant speech rose and fell more dramatically over the course of a chapter in these studies. In contrast, the two Round Robin studies showed the smallest amplitudes (e.g., Round Robin (Video Mediated) interactant $\alpha = 0.01$), suggesting relatively flat similarity trajectories with little temporal variation.

Target speech amplitudes were generally smaller than interactant speech amplitudes across studies, with STEM being a minor exception (target $\alpha = 0.03$ vs. interactant $\alpha = 0.02$). Length-scale estimates were more variable, with several studies (e.g., STEM, Negotiation, and the two Round Robin studies) producing credible intervals that spanned almost the entire prior. Where length-scales were more precisely estimated, the standard stimulus studies, New Mothers (interactant $\ell = 0.04$) and Middle Eastern Men (interactant $\ell = 0.05$), the length-scale values suggested that similarity was correlated across roughly 4–5% of a chapter before becoming effectively independent.

**Table 9:**

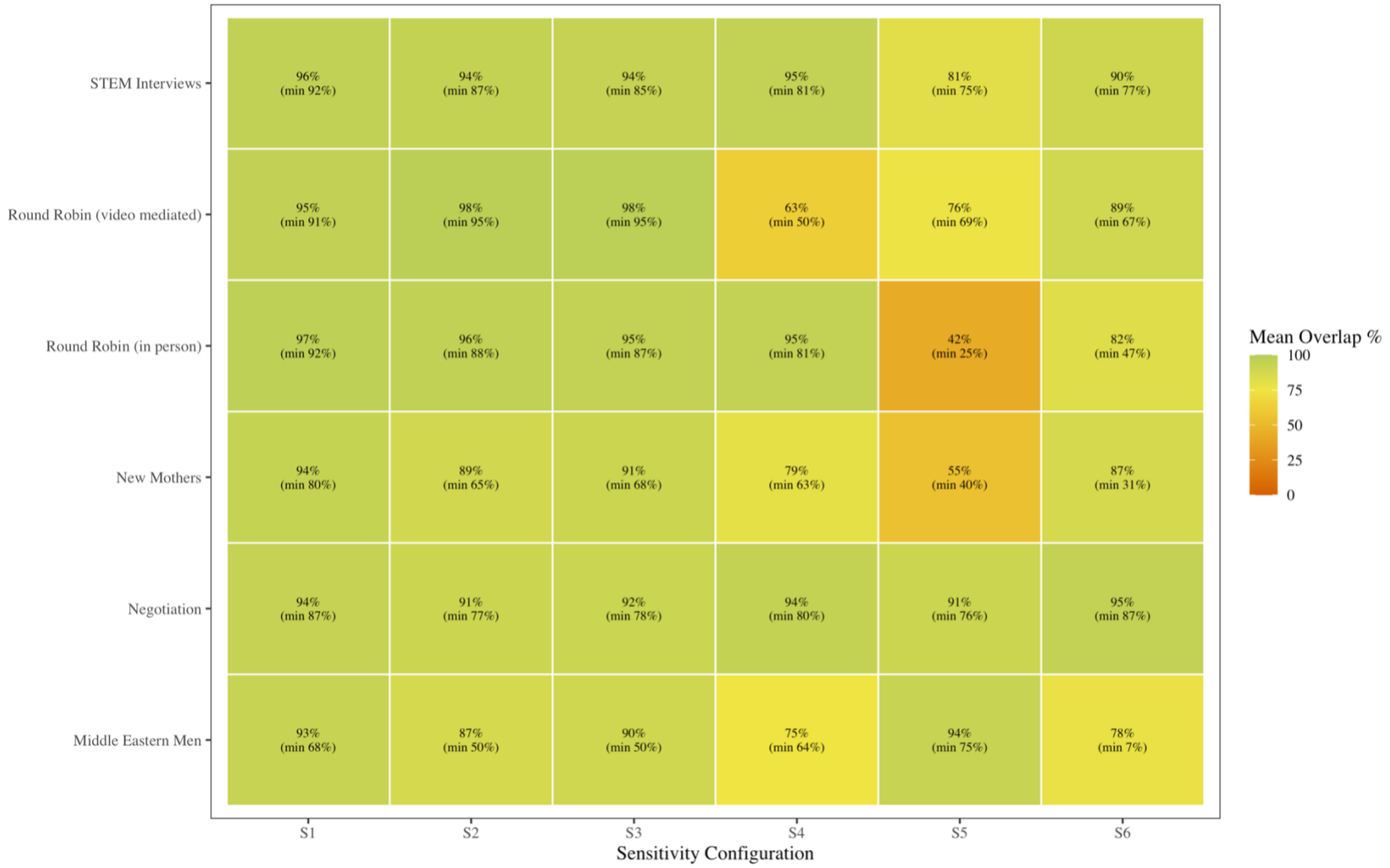
*Study 1: Individual Model Gaussian Process Hyperparameter Estimates with 95% CIs*

| Study | Variable | Mean | Median | SD | 95% CI |
|---|---|---|---|---|---|
| Middle Eastern Men | Amplitude (interactant) | 0.39 | 0.38 | 0.08 | 0.26, 0.59 |
| | Amplitude (target) | 0.09 | 0.09 | 0.03 | 0.05, 0.17 |
| | Length-scale (interactant) | 0.05 | 0.05 | 0.02 | 0.02, 0.10 |
| | Length-scale (target) | 0.05 | 0.04 | 0.03 | 0.01, 0.13 |
| Negotiation | Amplitude (interactant) | 0.06 | 0.05 | 0.03 | 0.02, 0.15 |
| | Amplitude (target) | 0.05 | 0.05 | 0.03 | 0.02, 0.13 |
| | Length-scale (interactant) | 0.78 | 0.64 | 0.61 | 0.04, 2.39 |
| | Length-scale (target) | 0.38 | 0.25 | 0.42 | 0.02, 1.48 |
| New Mothers | Amplitude (interactant) | 0.26 | 0.25 | 0.06 | 0.17, 0.40 |
| | Amplitude (target) | 0.05 | 0.04 | 0.02 | 0.02, 0.11 |
| | Length-scale (interactant) | 0.04 | 0.04 | 0.02 | 0.02, 0.09 |
| | Length-scale (target) | 0.29 | 0.26 | 0.19 | 0.03, 0.77 |
| Round Robin (In Person) | Amplitude (interactant) | 0.03 | 0.02 | 0.02 | 0.01, 0.09 |
| | Amplitude (target) | 0.00 | 0.00 | 0.01 | 0.00, 0.02 |
| | Length-scale (interactant) | 0.28 | 0.16 | 0.33 | 0.01, 1.18 |
| | Length-scale (target) | 0.40 | 0.05 | 2.02 | 0.01, 2.96 |
| Round Robin (Video Mediated) | Amplitude (interactant) | 0.01 | 0.00 | 0.01 | 0.00, 0.03 |
| | Amplitude (target) | 0.00 | 0.00 | 0.01 | 0.00, 0.02 |
| | Length-scale (interactant) | 0.38 | 0.06 | 1.04 | 0.01, 3.3 |
| | Length-scale (target) | 0.46 | 0.05 | 2.37 | 0.01, 3.85 |
| STEM | Amplitude (interactant) | 0.02 | 0.02 | 0.02 | 0.01, 0.07 |
| | Amplitude (target) | 0.03 | 0.03 | 0.03 | 0.01, 0.11 |
| | Length-scale (interactant) | 0.57 | 0.35 | 0.70 | 0.02, 2.52 |
| | Length-scale (target) | 0.58 | 0.43 | 0.54 | 0.03, 2.02 |

**Sensitivity Analysis Results.** To assess the robustness of findings to prior specification, all six individual-study models were re-estimated under each of the six alternative prior configurations (S1–S6) described above. Across all studies and all alternative prior sets, the substantive conclusions were generally unchanged: there did not seem to be a robust effect of speaker and there was a considerable amount of variability in the Gaussian process estimates. Figure 9 shows the mean posterior overlap between all six sets of sensitivity priors, with parameters varying on average by less than approximately 10% for the regularizing and diffuse priors S1-S3. The alternative Gaussian process kernel (S4) had poor model fit after several iterations of adjusting priors, suggesting that an over-smooth Gaussian process does not successfully explain the data in the inferential accuracy conversations.

**Figure 9:**

*Heatmap of Posterior Sensitivity Estimates*



Gaussian process hyperparameter estimates showed somewhat greater sensitivity to prior choice, particularly for the length-scale parameter in studies with fewer chapters (e.g., the Negotiation study). However, the direction and relative magnitude of speaker-type differences in both amplitude and length-scale remained consistent across specifications. The squared exponential kernel models (S4, S5) produced slightly smoother posterior trajectories than the Matérn 3/2 models, as expected, but the overall trajectory shapes and the relative positioning of target and interactant curves were comparable. The correlated random effects model (S6) yielded modest positive correlations between target and interactant perceiver-level random effects (ranging from 0.15 to 0.40 across studies), but the fixed effect and Gaussian process estimates were not meaningfully altered.

***Stage 2A: Univariate Meta-Analysis***

Following the meta-analytic specification described in the Method section, I pooled individual study estimates using Bayesian random-effects meta-analysis. Fixed effects and GP hyperparameters (log-transformed) were each meta-analyzed separately and are reported in Table 10. Pooled fixed effect estimates showed positive mean cosine similarity for both target and interactant speech across all six studies. Between-study heterogeneity ($\tau$) was modest, indicating reasonable consistency in mean-level semantic alignment across the different study designs. Between-study heterogeneity in the GP parameters was generally larger than for the fixed effects, reflecting the sensitivity of temporal dynamics to study-specific design features.

**Fixed Effects.** The pooled mean cosine similarity was 0.123 [95% CI: 0.10, 0.16] for target speech and 0.135 [95% CI: 0.09, 0.19] for interactant speech (see Figure 10), indicating no strong difference between the two. The between-study heterogeneity was small for target speech ($\tau$ = 0.018 [95% CI: 0.001, 0.075]) and somewhat larger for interactant speech ($\tau$ = 0.035 [95%

CI: 0.006, 0.127]), suggesting that the mean level of semantic similarity between interactant speech and perceiver inferences may vary slightly more across studies than the corresponding level for target speech, but, as the 95% credible intervals overlapped considerably, this difference was not substantial.

**Figure 10:**

*Study 1: Pooled Mean Semantic (Cosine) Similarity by Speaker and Study*

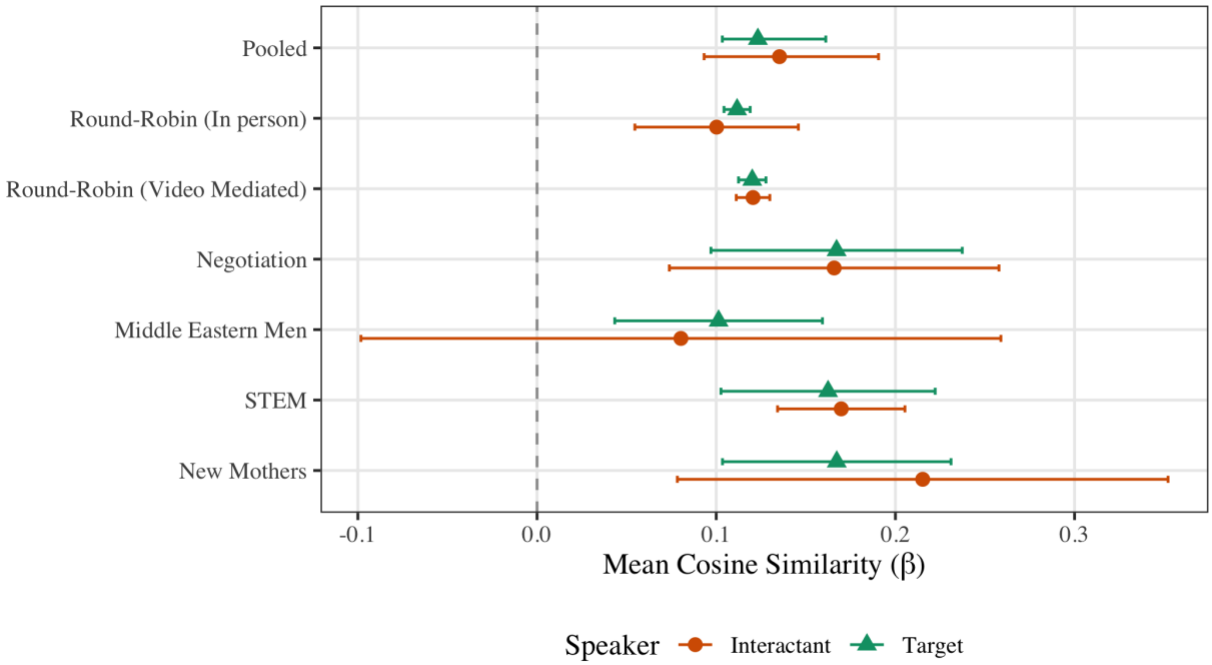


**Table 10:**

*Study 1: Stage 2A Univariate Meta-Analytic Pooled Parameter Estimates with 95% CIs*

| *Category* | *Parameter* | *Pooled* | *95% CI* | *τ [95% CI]* |
|---|---|---|---|---|
| Fixed Effects | β (Target) | 0.123 | [0.10, 0.16] | 0.018 [0.001, 0.075] |
| Fixed Effects | β (Interactant) | 0.135 | [0.09, 0.19] | 0.035 [0.006, 0.127] |
| GP Hyperparameters | α (Target) | 0.046 | [0.030, 0.069] | 0.077 [0.003, 0.246] |
| GP Hyperparameters | α (Interactant) | 0.092 | [0.054, 0.146] | 0.397 [0.262, 0.534] |
| GP Hyperparameters | ℓ (Target) | 0.134 | [0.089, 0.201] | 0.070 [0.003, 0.228] |
| GP Hyperparameters | ℓ (Interactant) | 0.101 | [0.070, 0.145] | 0.080 [0.004, 0.257] |

*Fixed effects on original scale; variance components back-transformed from log scale.*
*τ = between-study heterogeneity SD.*

**Gaussian Process Amplitude.** The pooled amplitude estimates provide information about the magnitude of temporal fluctuations in semantic similarity across a chapter. The pooled amplitude for target speech was 0.046 [95% CI: 0.030, 0.069], indicating that the Gaussian process function for target speech showed cosine similarity deviating from the mean by approximately ±0.046 across the typical chapter. The pooled amplitude for interactant speech was somewhat larger at 0.092 [95% CI: 0.054, 0.146], suggesting that semantic similarity between interactant speech and perceiver inferences showed somewhat larger temporal variability within chapters. However, the between-study heterogeneity for amplitude was substantial for both speaker types (target $\tau = 0.077$ [95% CI: 0.003, 0.246]; interactant $\tau = 0.397$ [95% CI: 0.262, 0.534]), indicating that the magnitude of temporal fluctuations varied

considerably across studies (particularly for interactant speech). This heterogeneity is explored further when the standard stimulus paradigm and dyadic interaction paradigm are compared but suggests that different studies produce different temporal dynamics in how speech content relates to perceiver inferences. The log-transformed pooled and individual study amplitudes are shown in Figure 11[24].

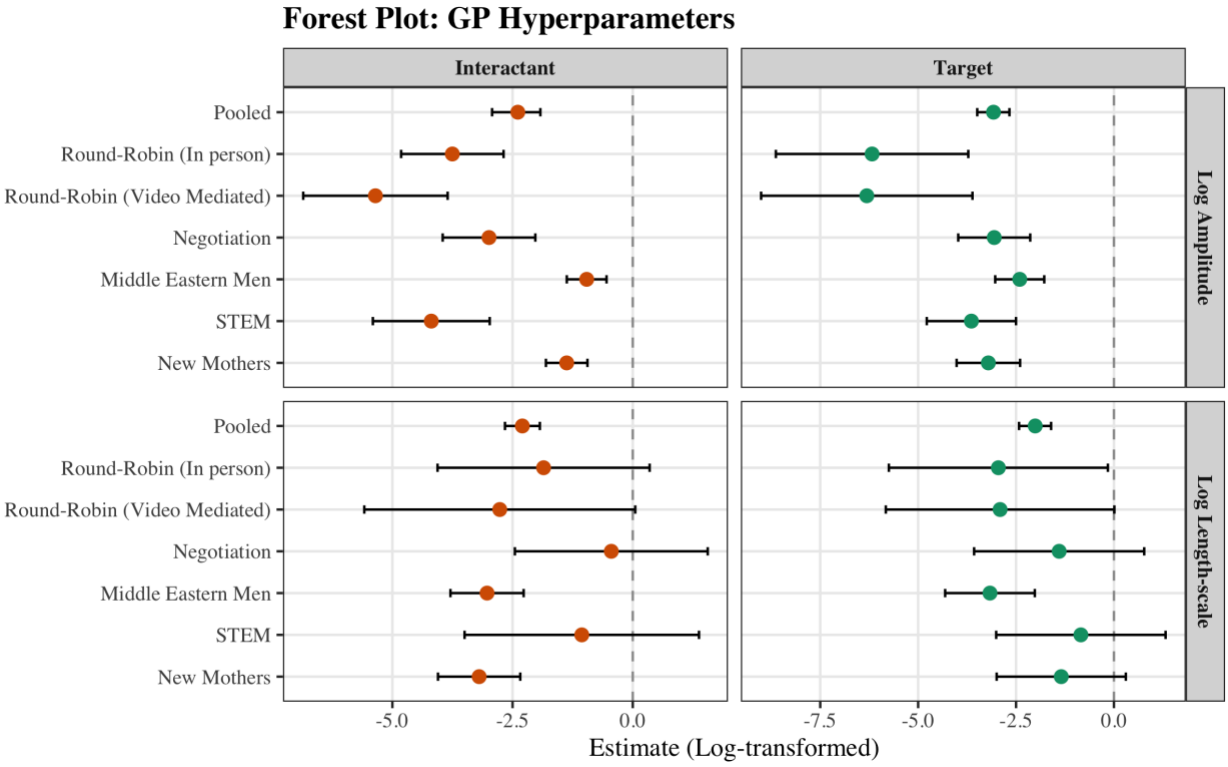
**Gaussian Process Length-Scale.** The pooled length-scale estimates characterize the temporal smoothness of the semantic similarity function. The pooled length-scale for target speech was 0.134 [95% CI: 0.089, 0.201], suggesting that, across studies, the target speech similarity function became substantially decorrelated after approximately 13% of the chapter had elapsed—or, in practical terms, that roughly 7 unique conversational moments occurred within a typical chapter of target speech. The pooled length-scale for interactant speech was similar at 0.101 [95% CI: 0.070, 0.145]; both are shown log-transformed with the individual study length-scale estimates in Figure 11 following the standard convention for reporting strictly positive Gaussian process hyperparameters (Rasmussen & Williams, 2008). Because length-scale parameters are bounded at zero and their posterior distributions tend to be right-skewed on the natural scale, log transformation yields a more symmetric distribution that is better suited for both optimization and meta-analytic pooling (e.g., Riutort-Mayol et al., 2023). Because all estimated length-scales fall between 0 and 1 (reflecting that the similarity function decorrelates within the span of a single chapter rather than over its full duration) their log-transformed values are necessarily negative. Between-study heterogeneity was moderate for both speaker types ($\tau$ =

[24] Log-transformed amplitude and length-scale estimates and 95% CIs are shown as the highly skewed data were un-readable without the transformation.

0.070 [95% CI: 0.003, 0.228] for target; τ = 0.080 [95% CI: 0.004, 0.257] for interactant), with comparable magnitudes across speaker types.

**Figure 11:**

*Study 1: Pooled and Individual Log-Transformed GP Amplitude and Length-Scale Estimates*




In examining these results, the individual-study posterior trajectories shown in Figure 8 provide a visual complement to these pooled estimates. In the standard stimulus paradigm studies (New Mothers, Middle Eastern Men), interactant speech showed notably steeper increases in semantic similarity toward the end of chapters compared to target speech, with wider credible bands reflecting greater uncertainty. In the dyadic interaction studies, the trajectories for target and interactant speech were more similar in shape and magnitude, with generally flatter trajectories and narrower credible intervals reflecting both the larger sample sizes in some of

these studies and the structural similarity of target and interactant roles in non-interview interactions.

**Between-Study Heterogeneity.** Figure 12 displays the full posterior distributions of $\tau$ (the between-study standard deviation from the random-effects meta-analysis) for all pooled parameters, with median and 95% credible interval quantile lines overlaid. The overall pattern reveals a clear hierarchy in how consistently different aspects of the semantic similarity model generalized across the six studies.

Fixed effect heterogeneity was relatively small: $\tau = 0.018$ [95% CI: 0.001, 0.075] for target speech and $\tau = 0.035$ [95% CI: 0.006, 0.127] for interactant[25]. Relative to the pooled estimates themselves ($\beta = 0.123$ and $\beta = 0.135$, respectively), these values indicate that the mean level of semantic similarity between speaker utterances and perceiver inferences was reasonably consistent across studies. The somewhat larger heterogeneity for interactant speech suggests that interactant-perceiver similarity may vary slightly more across study contexts, but as the 95% credible intervals overlap considerably, this difference does not appear sufficiently robust for further comment.

Heterogeneity in the Gaussian process hyperparameters was more pronounced, consistent with the expectation that temporal dynamics are more sensitive to study-specific design features than mean-level effects. The most striking value in Figure 12 is the interactant amplitude heterogeneity ($\tau = 0.397$ [95% CI: 0.262, 0.534]), which was substantially larger than the corresponding target value ($\tau = 0.077$ [95% CI: 0.003, 0.246]). In concrete terms, this suggests that the magnitude of temporal fluctuation in perceiver-interactant semantic similarity varied

[25] $\tau$ values in this section are reported on the log scale for GP hyperparameters; see Table 10 for back-transformed estimates on the original scale

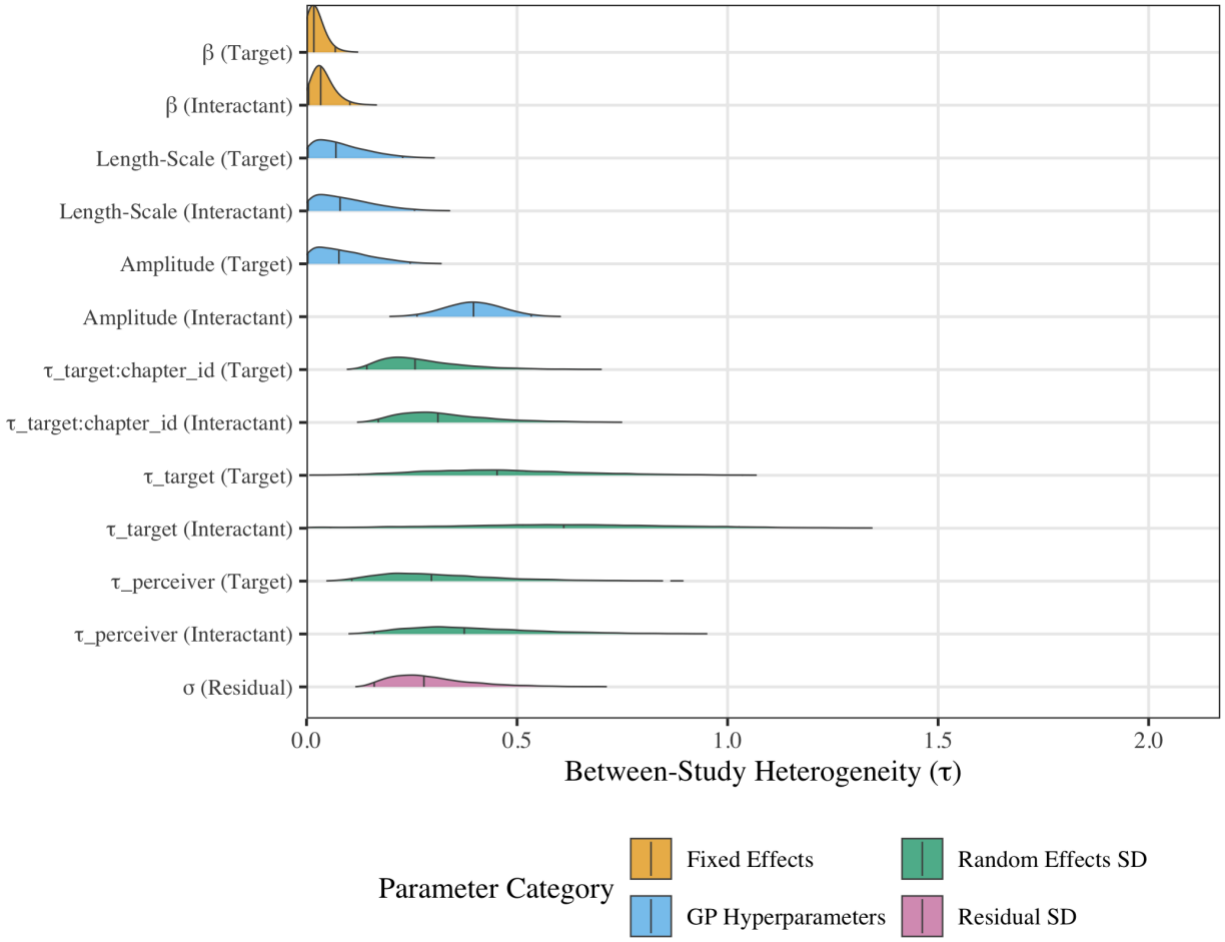
considerably across studies. Some studies produced large within-chapter swings in how interactant speech related to perceiver inferences, while others produced almost none. This asymmetry between speakers is explored further in the paradigm moderation analyses following the description of Hypothesis 4, but the amplitude $\tau$ alone strongly suggests that something about the interactant role differs across study designs in a way that affects temporal dynamics. Length-scale heterogeneity was moderate for both speaker types (target $\tau = 0.070$ [95% CI: 0.003, 0.228]; interactant $\tau = 0.080$ [95% CI: 0.004, 0.257]), indicating that the temporal smoothness of the similarity function also varied across studies, though to a lesser degree.

Random effect and residual variance components showed the largest between-study heterogeneity, with $\tau$ values ranging from 0.258 to 0.611 for the chapter-level random effects and $\tau = 0.279$ [95% CI: 0.161, 0.586] for the residual standard deviation. This is not surprising: these parameters capture study-specific differences in how variability is partitioned across perceivers, targets, and chapters, and the six studies differ considerably in sample size and conversation structure. The key point is that while the signal (fixed effects) was consistent across studies, the noise structure (variance components) was not, a pattern that may reflect the genuine diversity of paradigms and samples in these studies.

The substantial heterogeneity in interactant temporal dynamics captured by the large $\tau$ for interactant amplitude turns out to be one of the more theoretically meaningful findings from the meta-analysis. As will become apparent in the paradigm moderation analyses after Hypothesis 4, much of this heterogeneity is attributable to the standard stimulus/dyadic interaction contrast, in which interactants occupy fundamentally different conversational roles. I return to this point in Stage 2B and in the General Discussion.

**Figure 12:**

*Study 1: Posterior Distributions of Between-Study Heterogeneity (τ) for All Meta-Analytic Parameters*



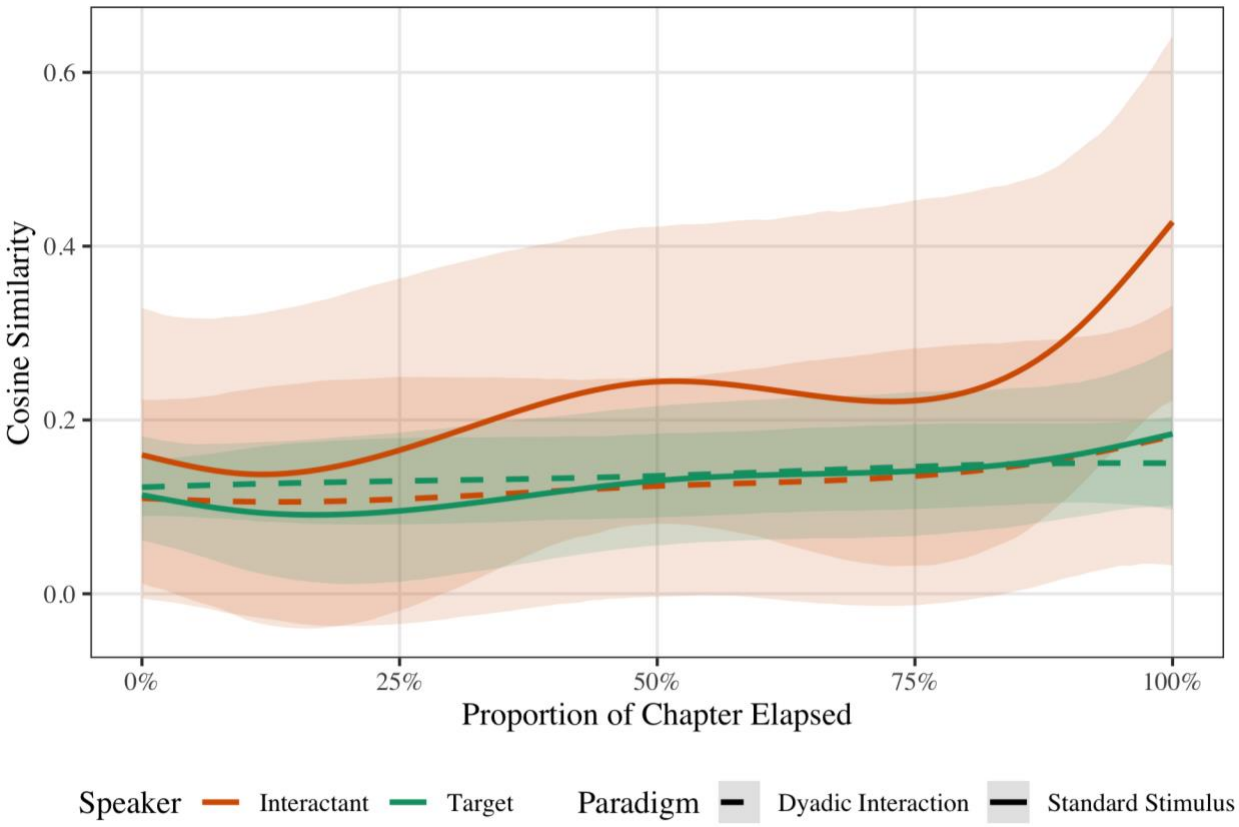

***Stage 2B: Multivariate Synthesis of Pointwise Posterior Trajectories***

Using the spline-based multivariate synthesis approach described in the Method section, I pooled pointwise posterior trajectories across the six studies to address Hypotheses 1–4. The same knot configuration (3 knots at 0.1, 0.5, 0.9) and basis coefficient structure were used throughout. Figure 13 displays the resulting pooled temporal trajectory of semantic similarity by speaker type and paradigm, with 95% credible intervals. The overall differences are reported here and the key paradigmatic differences are discussed below. The synthesized trajectory

reveals a gradual upward trend in semantic similarity over normalized chapter time, consistent with the prediction that perceiver inferences become more semantically aligned with the speech stream as utterances occur closer to the moment of inference.

**Figure 13:**

*Study 1: Pooled Temporal Trajectories of Semantic Similarity by Speaker Type and Paradigm with 95% CIs*




***Hypothesis Tests***

**Hypothesis 1: Temporal Increase.** *Prediction*: Semantic similarity between perceiver inferences and the speech stream will increase as utterances occur closer to the moment of inference. *Evidence*: Both operationalizations of temporal increase produced credible intervals excluding zero, with very strong evidence ratios in favor of a positive trend, providing consistent

evidence for a temporal increase in semantic alignment across the chapter: endpoint change (Δ = 0.109, 95% CI [0.056, 0.164], P(+) = 100%, ER > 1000), and median derivative (0.093, 95% CI [0.043, 0.160], P(+) = 99.9%, ER = 999). *Verdict*: Hypothesis 1 was supported (ER > 999 across methods). Results are presented separately for target and interactant speech in Table 11.

**Hypothesis 2: Speaker Differences.** *Prediction*: Target speech will show higher overall speech–inference similarity than interactant speech, reflecting a preferential focus on the person whose thoughts were being inferred. *Evidence*: The pooled mean cosine similarity was β = 0.123, 95% CI [0.10, 0.16] for target speech and β = 0.135, 95% CI [0.09, 0.19] for interactant speech. The synthesized trajectory contrast (Target − Interactant) for mean cosine similarity was −0.042, 95% CI [−0.175, 0.079], with Pr(Direction) = .233 and an evidence ratio of 0.30, indicating weak evidence in the direction opposite to the predicted target advantage. The interactant posterior was considerably more diffuse, reflecting roughly double the between-study heterogeneity (τ = 0.034 vs. 0.017 for target). *Verdict*: Hypothesis 2 was not supported. Perceiver inferences did not systematically show higher similarity with the target's verbal speech over the other's; the mean-level contrast was centered near zero with a wide credible interval spanning both directions.

**Table 11:**

*Study 1: Stage 2B Pooled Temporal Change Hypothesis Tests with 95% CIs*

| Paradigm | Speaker | Method | Estimate | 95% CI | Pr(Direction) | Evidence Ratio |
|---|---|---|---|---|---|---|
| Dyadic Interaction | Both | Endpoint (Δ t=0 to t=1) | 0.050 | [-0.003, 0.107] | 0.968 | 30.37 |
| | | Median Derivative | 0.046 | [-0.009, 0.120] | 0.947 | 17.91 |
| | Interactant | Endpoint (Δ t=0 to t=1) | 0.072 | [-0.024, 0.180] | 0.929 | 13.11 |
| | | Median Derivative | 0.062 | [-0.042, 0.208] | 0.883 | 7.52 |
| | Target | Endpoint (Δ t=0 to t=1) | 0.028 | [-0.013, 0.072] | 0.925 | 12.31 |
| | | Median Derivative | 0.030 | [-0.016, 0.082] | 0.921 | 11.70 |
| Overall | Both | Endpoint (Δ t=0 to t=1) | 0.109 | [0.056, 0.164] | >0.999 | >1000 |
| | | Median Derivative | 0.093 | [0.043, 0.160] | 0.999 | 999.00 |
| | Interactant | Endpoint (Δ t=0 to t=1) | 0.170 | [0.071, 0.271] | >0.999 | >1000 |
| | | Median Derivative | 0.137 | [0.052, 0.256] | 0.997 | 284.71 |
| | Target | Endpoint (Δ t=0 to t=1) | 0.049 | [0.011, 0.095] | 0.993 | 139.35 |
| | | Median Derivative | 0.051 | [0.004, 0.103] | 0.981 | 52.33 |
| Standard Stimulus | Both | Endpoint (Δ t=0 to t=1) | 0.169 | [0.089, 0.250] | >0.999 | >1000 |
| | | Median Derivative | 0.143 | [0.069, 0.231] | >0.999 | >1000 |
| | Interactant | Endpoint (Δ t=0 to t=1) | 0.268 | [0.121, 0.412] | 0.999 | 999.00 |
| | | Median Derivative | 0.216 | [0.095, 0.373] | 0.997 | 379.95 |
| | Target | Endpoint (Δ t=0 to t=1) | 0.070 | [0.009, 0.142] | 0.984 | 60.54 |
| | | Median Derivative | 0.072 | [-0.006, 0.158] | 0.969 | 31.39 |

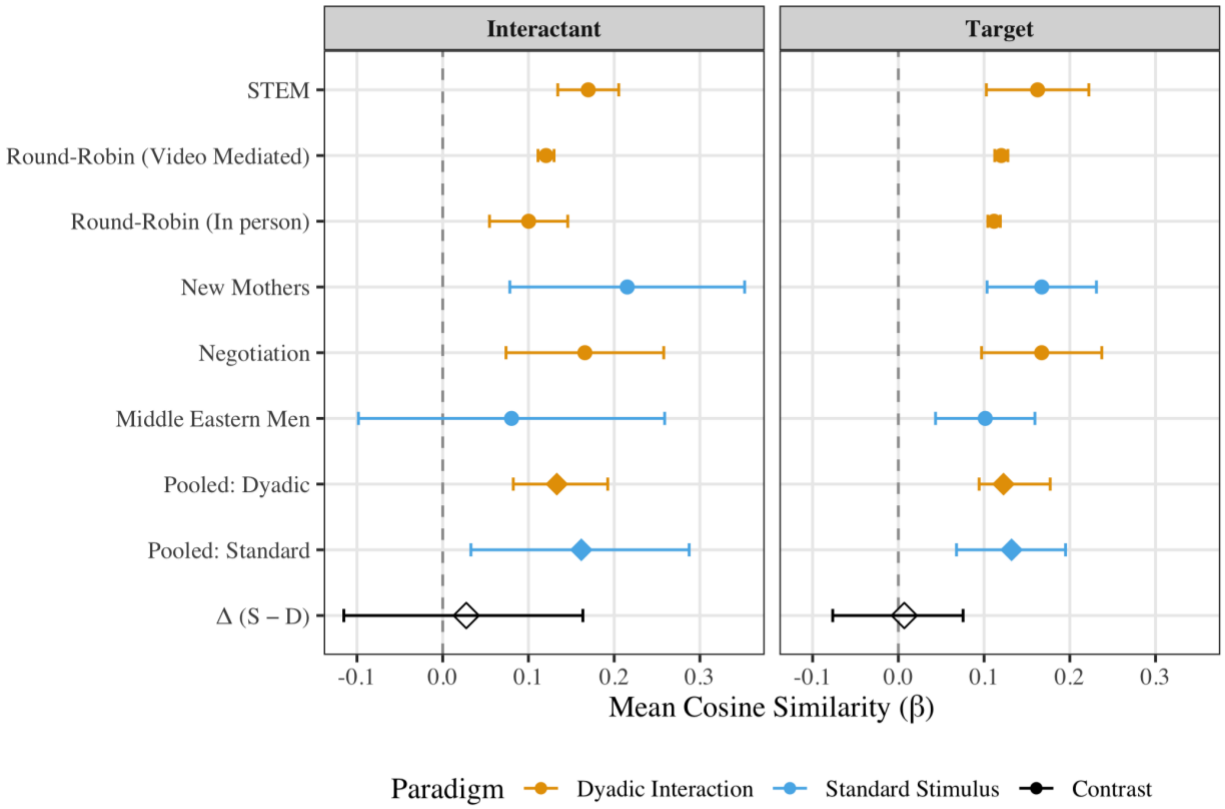
**Hypothesis 3: Differential Rates of Temporal Change.** *Prediction:* Speech–inference similarity will increase closer to the inference point for both speakers, but target speech will show a stronger increase; in other words, although all utterances were expected to increase in semantic similarity closer to the inference point, I expected target speech to show a larger increase closer to the inference point (i.e., a combination of Hypotheses 1 and 2). *Evidence*: Both speakers showed positive temporal trends (endpoint $\Delta = 0.049$ for target, ER = 139.35; $\Delta = 0.170$ for interactant, ER > 1000), however, contrary to the prediction*, interactant speech showed, if anything, a substantially larger* increase in semantic similarity. The speaker contrast in temporal change was −0.121, 95% CI [−0.227, −0.013], with Probability(Target > Interactant) = .016 and an evidence ratio of 0.02, very strong evidence *against* the predicted target preference. This was compounded by the finding that the interactant posterior was considerably more diffuse, reflecting greater between-study heterogeneity in interactant temporal dynamics. *Verdict*: Hypothesis 3 was not supported. Both speakers showed strong evidence for increased semantic similarity closer to the inference, but the predicted target advantage in rate of change was not observed; the pattern was, if anything, reversed.

**Hypothesis 4: Paradigm Moderation.** *Prediction*: Interactant speech in the dyadic interaction paradigm will be more semantically similar to perceiver inferences than interactant speech in the standard stimulus paradigm, reflecting perceivers' preferential attention to their own speech when they are the interactant. *Evidence*: The pooled mean cosine similarity for interactant speech was numerically higher in the standard stimulus studies than in the dyadic interaction studies (Figure 14), the opposite direction from what was predicted. The paradigm moderator explained none of the between-study heterogeneity in interactant mean similarity ($R^2\_\tau = 0.0\%$, 95% CI [0.0%, 98.9%]), and the leave-one-study-out cross-validation did not

favor the paradigm-moderated model over the unconditional model (ΔELPD = −1.10, SE = 0.56). *Verdict*: Hypothesis 4 was not supported. Perceivers did not show preferential weighting of their own speech in dyadic interaction paradigm studies; if anything, interactant speech was slightly *more* aligned with perceiver inferences in the standard stimulus paradigm, where the interactant was *not* the perceiver. The wide $R^2\_\tau$ credible interval reflects the limited precision available with six studies for detecting study-level moderation of mean similarity.

**Figure 14:**

*Study 1: Individual and Pooled Paradigm Differences in Cosine Similarity*



***Exploratory Analysis: Digging Deeper into Paradigm Differences.***

Although Hypothesis 4 tested a specific directional prediction about mean-level interactant similarity, the paradigm distinction between standard stimulus and dyadic interaction studies proved to be a far more consequential moderator of *temporal* dynamics than of overall similarity levels. The following exploratory analyses examine paradigm effects across the full set of model parameters.

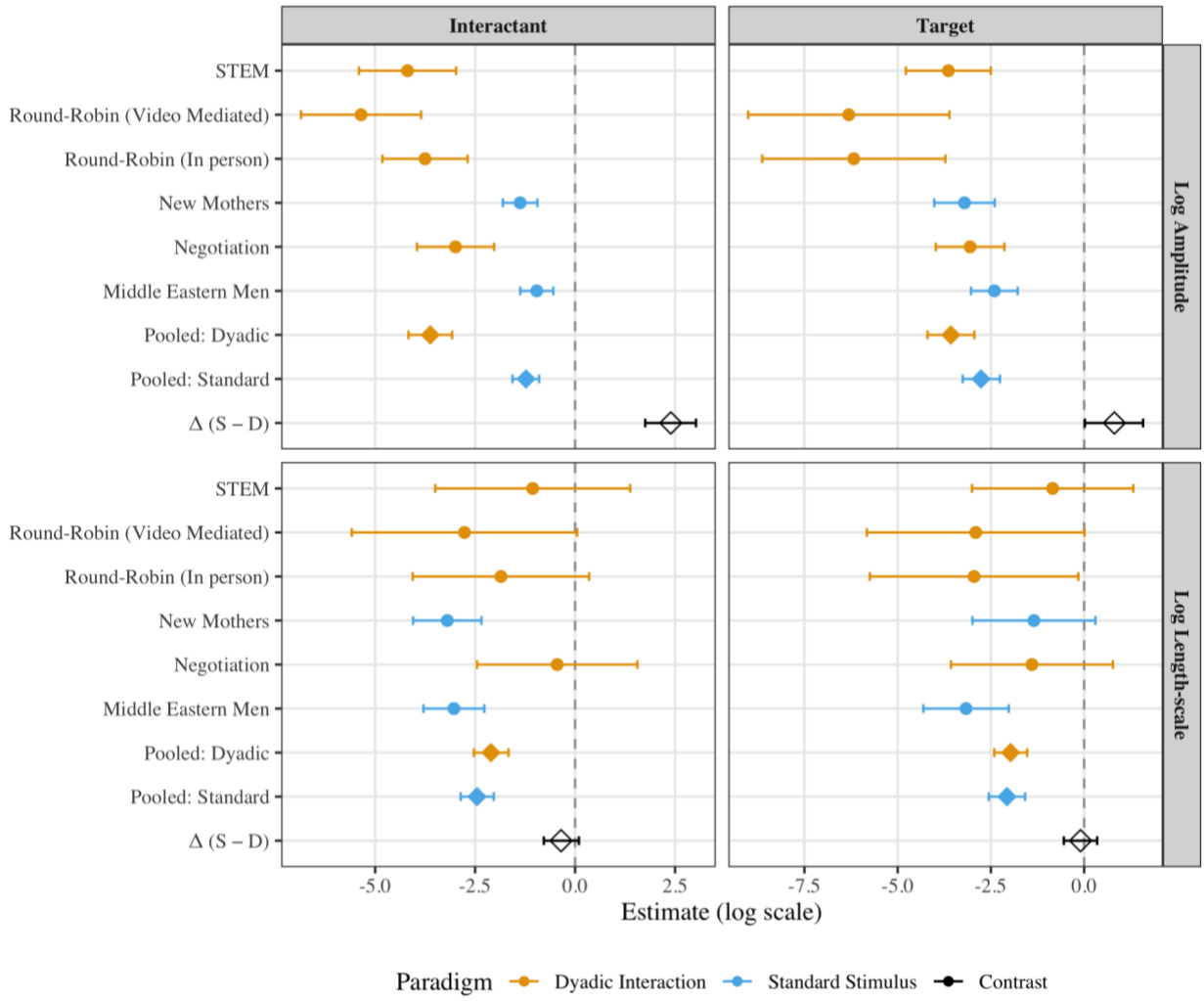
**Temporal Change by Paradigm.** The temporal increase in speech–inference similarity (i.e., Hypothesis 1) was decisively supported in the standard stimulus studies (endpoint Δ = 0.169 [0.089, 0.250], P(+) = 100%, ER > 1000) but more uncertain in dyadic interaction studies (endpoint Δ = 0.050 [−0.003, 0.107], P(+) = 96.8%, ER = 30.37). The overall paradigm contrast in temporal change was credible (Dyadic − Standard endpoint = −0.120 [−0.204, −0.031]), indicating that standard stimulus studies produced reliably stronger temporal effects.

**Speaker-Specific Temporal Effects**. The paradigm difference was driven primarily by interactant speech. Within standard stimulus studies, interactant speech showed the strongest temporal effect (endpoint Δ = 0.268 [0.121, 0.412], ER = 999), whereas target speech showed a smaller but reliable increase (endpoint Δ = 0.070 [0.009, 0.142], ER = 60.54). The paradigm contrast for interactant temporal change was large and credible (Dyadic − Standard = −0.197 [−0.345, −0.038], ER = 0.01), but the corresponding contrast for target speech was smaller and not credible (−0.043 [−0.122, 0.028], ER = 0.13). In the standard stimulus paradigm, the speaker contrast in temporal change credibly favored the interactant (Target − Interactant endpoint = −0.198 [−0.353, −0.034], ER = 0.01); in dyadic interaction studies, no speaker difference was observed (−0.044 [−0.161, 0.062], ER = 0.26).

**Gaussian Process Hyperparameters**. The paradigm distinction was clearly reflected in the GP hyperparameter estimates for interactant speech (Figure 15). In standard stimulus paradigm studies, interactant speech showed higher amplitude (larger temporal fluctuations in similarity) and shorter length-scales (more rapid change) than in dyadic interaction paradigm studies, with the amplitude contrast excluding zero. For target speech, GP hyperparameters did not differ meaningfully by paradigm. These estimates are interpretable in terms of the interviewer role occupied by interactants in standard stimulus studies: interviewer utterances are less frequent and more variable in their informational value, producing an intermittent pattern of high and low similarity that the GP captures as high amplitude with a short length-scale.

**Figure 15:**

*Study 1: Individual and Pooled Paradigm Differences in GP Hyperparameter Estimates (Log-Transformed)*




**Between-Study Heterogeneity Decomposition**. The paradigm moderation analyses established that the standard stimulus/dyadic interaction contrast moderated several meta-analytic parameters but left open the question of *how much* of the between-study heterogeneity was attributable to this distinction. To address this, I computed a Bayesian $R^2$-analog for $\tau$ (Table 12). For each meta-analytic parameter, I extracted the paired posterior draws of $\tau$ from the unconditional (Stage 2A) and paradigm-moderated models. I then computed, at each posterior

draw, $R^2_\tau = 1 - (\tau^2_{moderated} / \tau^2_{unconditional})$, giving a full posterior distribution of the proportion of between-study variance explained by paradigm, rather than a single point estimate.

**Heterogeneity Decomposition.** Paradigm type explained a very large amount of between-study heterogeneity in several interactant-specific parameters: interactant GP amplitude ($R^2_\tau$ = 96.4% [58.3%, 100.0%]), perceiver-level random effects variance ($R^2_\tau$ = 86.3% [0.0%, 100.0%]), and target-level random effects variance ($R^2_\tau$ = 85.0% [0.0%, 100.0%]). By contrast, paradigm explained little of the between-study variance in mean similarity for either speaker ($R^2_\tau$ = 0.0% for both β parameters) or in target-specific GP hyperparameters ($R^2_\tau \leq 13\%$). This pattern supports the conclusion that the diffuse interactant posteriors observed under Hypothesis 3 were largely attributable to the standard stimulus/dyadic interaction distinction and that paradigm moderation selectively affected the *temporal structure* of interactant speech rather than its overall similarity level.

**Table 12:**

*Study 1: Between-Study Heterogeneity Decomposition by Paradigm Type*

| Category | Parameter | τ (Unconditional) | τ (Paradigm-Moderated) | $R^2\_\tau$ [95% CI] |
|---|---|---|---|---|
| Fixed Effects | β (Interactant) | 0.034 | 0.035 | 0.0% [0.0%, 98.9%] |
| Fixed Effects | β (Target) | 0.017 | 0.021 | 0.0% [0.0%, 99.6%] |
| GP Hyperparameters | α (Interactant) | 0.397 | 0.074 | 96.4% [58.3%, 100.0%] |
| GP Hyperparameters | α (Target) | 0.077 | 0.073 | 8.2% [0.0%, 99.8%] |
| GP Hyperparameters | ℓ (Interactant) | 0.080 | 0.074 | 13.0% [0.0%, 99.9%] |
| GP Hyperparameters | ℓ (Target) | 0.070 | 0.070 | 4.2% [0.0%, 99.8%] |
| Random Effects SD | τ_perceiver (Interactant) | 0.375 | 0.141 | 86.3% [0.0%, 100.0%] |
| Random Effects SD | τ_perceiver (Target) | 0.297 | 0.351 | 0.0% [0.0%, 90.5%] |
| Random Effects SD | τ_target (Interactant) | 0.611 | 0.222 | 85.0% [0.0%, 100.0%] |
| Random Effects SD | τ_target (Target) | 0.452 | 0.243 | 71.8% [0.0%, 100.0%] |
| Random Effects SD | τ_target:chapter_id (Interactant) | 0.312 | 0.342 | 0.0% [0.0%, 84.7%] |
| Random Effects SD | τ_target:chapter_id (Target) | 0.258 | 0.264 | 0.0% [0.0%, 86.4%] |
| Residual SD | σ (Residual) | 0.279 | 0.298 | 0.0% [0.0%, 83.3%] |

*$R^2\_\tau = 1 - (\tau^2\_\text{moderated} / \tau^2\_\text{unconditional})$. Values represent the posterior median [95% CI] of the proportion of between-study variance explained by the standard stimulus vs. dyadic interaction contrast. Bounded at 0%.*

The decomposition confirmed what the τ ridge plot (Figure 12) suggested: paradigm explained the most heterogeneity for GP hyperparameters, particularly interactant amplitude. For

interactant amplitude, which had the largest unconditional τ, a substantial proportion of the between-study variance was attributable to the paradigm distinction. This is consistent with the argument developed in the paradigm moderation section: the elevated temporal variability for interactant speech in standard stimulus paradigm studies likely reflects the distinctive interviewer role, and once this structural difference is accounted for, the remaining between-study variability is substantially reduced.

For the fixed effects, the $R^2\_\tau$ posteriors were wide and centered near modest values, consistent with the small unconditional τ values for these parameters. When there is little heterogeneity to begin with, paradigm cannot explain much of it, and the decomposition is thus poorly identified. For the variance components, $R^2\_\tau$ values were variable, reflecting the fact that the variance structure of individual studies depends on many features beyond paradigm type (sample size, nesting depth, conversation length) that are not captured by a binary moderator.

***Cross-Validation***. As a complementary assessment, I compared the unconditional and paradigm-moderated models using approximate leave-one-study-out cross-validation (LOO-CV). Leave-one-study-out cross-validation strongly favored the paradigm-moderated model for interactant GP amplitude (ΔELPD = −15.03, SE = 3.94), with more modest support for interactant length-scale (ΔELPD = −1.68, SE = 0.72). The paradigm model was not favored for fixed effects (β Target: −1.22 (SE = 0.76); β Interactant: −1.10 (SE = 0.56)) or residual variance, consistent with paradigm moderation operating primarily through the temporal structure of interactant speech rather than mean similarity levels. The LOO-CV estimates carry high variance with only six studies, so these should be interpreted alongside the $R^2\_\tau$ decomposition rather than as standalone evidence.

To evaluate the reliability of the LOO-CV approximation, I examined Pareto k diagnostic values for each observation across all primary specification (S0) models. Pareto k values index the influence of individual observations on the posterior; values below 0.50 indicate reliable LOO estimates, values between 0.50 and 0.70 are acceptable, values between 0.70 and 1.00 are borderline, and values exceeding 1.00 indicate that the LOO approximation may be unreliable for that observation (Vehtari et al., 2021). Across all models, no observation produced a Pareto k value exceeding 1.00, and mean Pareto k values were uniformly low (all $\leq 0.112$), indicating that high-influence observations were isolated rather than systematic. At the borderline threshold ($k > 0.7$), 16 of the 34 LOO-CV-eligible models contained at least one such observation, with GP models more frequently affected than their linear counterparts, consistent with the greater flexibility of the Gaussian process specification. To accommodate borderline observations, moment matching was applied to all LOO-CV computations (Paananen et al., 2021), which adjusts the importance sampling weights for observations with $k > 0.70$ and improves the stability of the LOO-CV estimates without discarding data. The LOO-CV results (Table 13) were broadly consistent with the $R^2\_\tau$ decomposition: parameters where paradigm explained substantial heterogeneity tended to show $\Delta$ELPD values favoring the moderated model, though the uncertainty was considerable.

**Table 13:**

*Study 1: LOO-CV Model Comparison: Unconditional vs. Paradigm-Moderated Meta-Analyses*

| Parameter | ΔELPD (SE) | Favors Paradigm Model |
|---|---|---|
| β (Target) | -1.22 (0.76) | No |
| β (Interactant) | -1.10 (0.56) | No |
| α (Target) | -1.83 (2.84) | Yes |
| α (Interactant) | -15.03 (3.94) | Yes |
| ℓ (Target) | -0.02 (0.12) | No |
| ℓ (Interactant) | -1.68 (0.72) | Yes |
| σ (Residual) | -0.44 (0.18) | No |

*ΔELPD = difference in expected log-predictive density (negative values favor the paradigm-moderated model). With k = 6 studies, LOO-CV estimates have high variance; these should be interpreted alongside the $R^2_\tau$ decomposition rather than as standalone evidence.*

**Quintile-Level Decomposition of Temporal Change**. To examine where in the chapter the temporal effects are concentrated, I computed within-quintile median derivatives for each speaker and paradigm. This can be thought of as a descriptive rather than a meaningful statistical test: there is no theoretical reason for splitting a stimulus chapter into quintiles, but doing so may help describe where changes in semantic similarity are occurring. Results are presented in Figure 16 and Table 14. The quintile-level decomposition revealed that the temporal increase in semantic similarity was not uniform across the chapter, nor across paradigm.

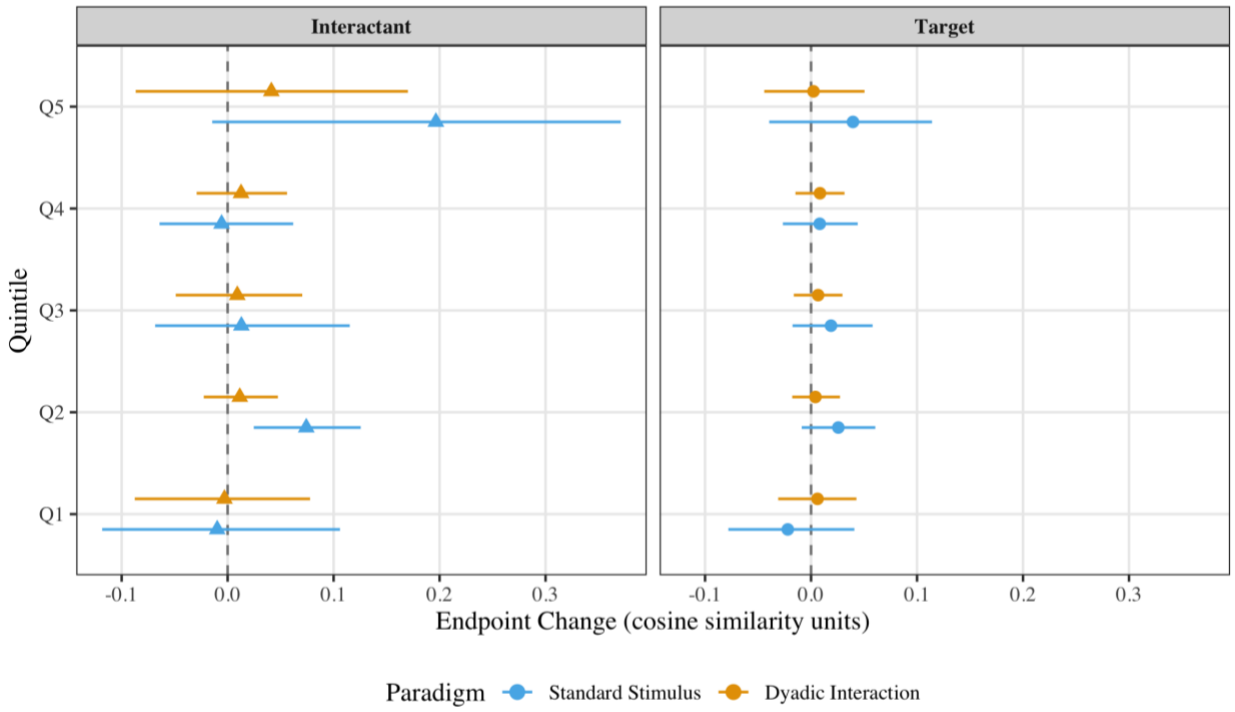
For the overall trajectory, successive quintile contrasts showed a gradual acceleration: the Q2–Q1 contrast was near zero (0.005 [−0.011, 0.021]), with progressively larger point estimates through Q3–Q2 (0.008 [−0.020, 0.037]), Q4–Q3 (0.009 [−0.020, 0.039]), and Q5–Q4 (0.016 [−0.017, 0.050]). This somewhat accelerating pattern (albeit hindered by the wide variability)

suggests that the temporal alignment between perceiver inferences and the speech stream strengthens in the latter portions of the chapter, particularly in the final quintile.

The variability was considerably reduced when split by paradigm. The temporal effect in standard stimulus interactant speech was not uniform across the chapter. In the quintile decomposition (Table 14), the strongest interactant effects in the standard stimulus paradigm appeared in Q2 (endpoint Δ = 0.074 [0.025, 0.126], ER = 443) and Q5 (endpoint Δ = 0.197 [−0.015, 0.371], ER = 29.65), with the median derivative reaching its largest value in Q5 (1.058, ER = 29.77). In the dyadic interaction paradigm, no individual quintile showed a credible interactant effect. This concentration of the standard stimulus effect in specific quintiles is consistent with the GP characterization: interviewer speech produced localized elevations in similarity at particular moments rather than a uniform trend.

**Figure 16:**

*Study 1: Quintile-Level Median Derivatives by Speaker and Paradigm with 95% CIs*



The speaker-level quintile analyses capture the overall pattern. For target speech, the temporal change was relatively stable across quintiles, consistent with a gradual, steady increase. In contrast, interactant speech showed a pronounced acceleration in the final quintile suggesting a recency effect in which the most recent interactant utterances exert the strongest influence on perceiver inferences. This recency effect was particularly striking in standard stimulus paradigm studies, where the Q5–Q4 interactant contrast was the largest segmented effect observed.

**Table 14:**

*Study 1: Quintile-Level Change by Speaker and Paradigm*

| Quintile | Measure | Paradigm | Speaker | Estimate | 95% CI | Pr(Direction) | Evidence Ratio |
|---|---|---|---|---|---|---|---|
| Q1 | Endpoint Change | Dyadic Interaction | Interactant | -0.003 | [−0.088, 0.078] | 0.474 | 0.900 |
| Q1 | Endpoint Change | Dyadic Interaction | Target | 0.006 | [−0.031, 0.043] | 0.653 | 1.880 |
| Q1 | Endpoint Change | Standard Stimulus | Interactant | -0.010 | [−0.118, 0.106] | 0.416 | 0.710 |
| Q1 | Endpoint Change | Standard Stimulus | Target | -0.022 | [−0.078, 0.041] | 0.199 | 0.250 |
| Q1 | Median Derivative | Dyadic Interaction | Interactant | -0.021 | [−0.487, 0.419] | 0.466 | 0.870 |
| Q1 | Median Derivative | Dyadic Interaction | Target | 0.031 | [−0.172, 0.232] | 0.643 | 1.81 |
| Q1 | Median Derivative | Standard Stimulus | Interactant | -0.089 | [−0.684, 0.548] | 0.375 | 0.60 |
| Q1 | Median Derivative | Standard Stimulus | Target | -0.129 | [−0.434, 0.213] | 0.182 | 0.22 |
| Q2 | Endpoint Change | Dyadic Interaction | Interactant | 0.011 | [−0.023, 0.047] | 0.751 | 3.01 |
| Q2 | Endpoint Change | Dyadic Interaction | Target | 0.004 | [−0.018, 0.027] | 0.661 | 1.95 |
| Q2 | Endpoint Change | Standard Stimulus | Interactant | 0.074 | [0.025, 0.126] | 0.998 | 443.44 |
| Q2 | Endpoint Change | Standard Stimulus | Target | 0.026 | [−0.009, 0.061] | 0.938 | 15.19 |
| Q2 | Median Derivative | Dyadic Interaction | Interactant | 0.059 | [−0.119, 0.249] | 0.750 | 3.01 |
| Q2 | Median Derivative | Dyadic Interaction | Target | 0.020 | [−0.095, 0.140] | 0.643 | 1.81 |
| Q2 | Median Derivative | Standard Stimulus | Interactant | 0.389 | [0.136, 0.664] | 0.998 | 420.05 |
| Q2 | Median Derivative | Standard Stimulus | Target | 0.136 | [−0.047, 0.320] | 0.938 | 15.00 |

| Quintile | Measure | Paradigm | Speaker | Estimate | 95% CI | Pr(Direction) | Evidence Ratio |
|---|---|---|---|---|---|---|---|
| Q3 | Endpoint Change | Dyadic Interaction | Interactant | 0.009 | [−0.049, 0.070] | 0.620 | 1.63 |
| Q3 | Endpoint Change | Dyadic Interaction | Target | 0.007 | [−0.016, 0.030] | 0.739 | 2.83 |
| Q3 | Endpoint Change | Standard Stimulus | Interactant | 0.013 | [−0.068, 0.115] | 0.596 | 1.47 |
| Q3 | Endpoint Change | Standard Stimulus | Target | 0.019 | [−0.017, 0.058] | 0.861 | 6.17 |
| Q3 | Median Derivative | Dyadic Interaction | Interactant | 0.047 | [−0.265, 0.376] | 0.615 | 1.60 |
| Q3 | Median Derivative | Dyadic Interaction | Target | 0.035 | [−0.090, 0.158] | 0.733 | 2.75 |
| Q3 | Median Derivative | Standard Stimulus | Interactant | 0.060 | [−0.381, 0.615] | 0.577 | 1.36 |
| Q3 | Median Derivative | Standard Stimulus | Target | 0.099 | [−0.098, 0.310] | 0.854 | 5.84 |
| Q4 | Endpoint Change | Dyadic Interaction | Interactant | 0.013 | [−0.029, 0.056] | 0.733 | 2.74 |
| Q4 | Endpoint Change | Dyadic Interaction | Target | 0.008 | [−0.015, 0.032] | 0.797 | 3.93 |
| Q4 | Endpoint Change | Standard Stimulus | Interactant | -0.006 | [−0.064, 0.062] | 0.407 | 0.69 |
| Q4 | Endpoint Change | Standard Stimulus | Target | 0.008 | [−0.027, 0.044] | 0.697 | 2.30 |
| Q4 | Median Derivative | Dyadic Interaction | Interactant | 0.059 | [−0.158, 0.282] | 0.710 | 2.45 |
| Q4 | Median Derivative | Dyadic Interaction | Target | 0.043 | [−0.078, 0.163] | 0.792 | 3.82 |
| Q4 | Median Derivative | Standard Stimulus | Interactant | -0.066 | [−0.368, 0.292] | 0.319 | 0.47 |
| Q4 | Median Derivative | Standard Stimulus | Target | 0.034 | [−0.147, 0.221] | 0.654 | 1.89 |
| Q5 | Endpoint Change | Dyadic Interaction | Interactant | 0.041 | [−0.087, 0.170] | 0.750 | 3.01 |
| Q5 | Endpoint Change | Dyadic Interaction | Target | 0.002 | [−0.044, 0.050] | 0.532 | 1.14 |

| Quintile | Measure | Paradigm | Speaker | Estimate | 95% CI | Pr(Direction) | Evidence Ratio |
|---|---|---|---|---|---|---|---|
| Q5 | Endpoint Change | Standard Stimulus | Interactant | 0.197 | [−0.015, 0.371] | 0.967 | 29.65 |
| Q5 | Endpoint Change | Standard Stimulus | Target | 0.040 | [−0.039, 0.114] | 0.866 | 6.46 |
| Q5 | Median Derivative | Dyadic Interaction | Interactant | 0.216 | [−0.473, 0.902] | 0.746 | 2.94 |
| Q5 | Median Derivative | Dyadic Interaction | Target | 0.009 | [−0.241, 0.266] | 0.521 | 1.09 |
| Q5 | Median Derivative | Standard Stimulus | Interactant | 1.058 | [−0.083, 1.996] | 0.968 | 29.77 |
| Q5 | Median Derivative | Standard Stimulus | Target | 0.210 | [−0.217, 0.606] | 0.866 | 6.43 |

*Endpoint Change = trajectory value at quintile end minus quintile start. Median Derivative = median instantaneous growth velocity within quintile. Pr(Direction) = posterior probability estimate > 0. Evidence Ratio = posterior odds.*

**Posterior-Mean Approximation Diagnostic.** The variance ratio diagnostic assessed whether the posterior-mean approximation used in the spline projection step is justified by comparing between-study heterogeneity to within-study posterior uncertainty for each spline coefficient. All 10 spline coefficients (5 per speaker) exceeded the threshold ratio of 3.0, with an overall mean ratio of 28.44 (Table 15). Target speech coefficients showed ratios ranging from 4.9 (β4) to 23.0 (β5), while interactant speech coefficients showed uniformly higher ratios ranging from 16.9 (β4) to 81.1 (β5). The substantially larger ratios for interactant speech reflect the greater between-study heterogeneity in interactant temporal dynamics, consistent with the wider credible intervals and larger amplitude estimates observed in the Stage 2A analyses. The universally high ratios confirm that between-study variability dominates within-study posterior uncertainty for all coefficients, providing strong justification for the posterior-mean approximation in this dataset.

**Table 15:**

*Study 1: Variance Ratio Diagnostics*

| Speaker | Coefficient | Within-Study Var | Between-Study Var | Ratio |
|---|---|---|---|---|
| Target | β1 | 0.000118 | 0.000862 | 7.30 |
| | β2 | 0.000095 | 0.001204 | 12.70 |
| | β3 | 0.000110 | 0.000994 | 9.00 |
| | β4 | 0.000387 | 0.001900 | 4.90 |
| | β5 | 0.000076 | 0.001759 | 23.00 |
| Interactant | β1 | 0.000328 | 0.007317 | 22.30 |
| | β2 | 0.000191 | 0.007728 | 40.40 |
| | β3 | 0.000287 | 0.019126 | 66.70 |
| | β4 | 0.000753 | 0.012744 | 16.90 |
| | β5 | 0.000520 | 0.042188 | 81.10 |

Although the trajectory synthesis provided pooled estimates across the six studies, the variance ratio diagnostic and the between-study heterogeneity estimates from Stage 2A suggest that the temporal dynamics of semantic similarity vary meaningfully across studies. The between-study heterogeneity standard deviations for interactant amplitude ($\tau = 0.397$ [0.262, 0.534]) and length-scale ($\tau = 0.080$ [0.004, 0.257]) were substantial relative to the pooled parameter estimates, indicating that the degree of temporal fluctuation in semantic similarity differs substantially from one study context to the next. This heterogeneity is further evidenced by the variance ratio diagnostic, in which interactant coefficients showed ratios up to 81.1 (an order of magnitude larger than the minimum threshold) reflecting the fact that studies differed substantially in how interactant speech aligned with perceiver inferences over time. The paradigm moderation results (H4) account for a portion of this variability, but the residual between-study heterogeneity suggests that other study-level factors (e.g., conversational topic,

relationship context, recording modality) also contribute to the diverse temporal patterns observed across these six inferential accuracy studies.

## Discussion

### *Confirmatory Hypotheses*

The four hypotheses I set out to address were: whether speech–inference similarity increases over the course of a conversation (H1), whether target speech is more similar to perceiver inferences than interactant speech (H2), whether target speech shows a larger temporal increase than interactant speech (H3), and whether interactant speech in dyadic interaction paradigm studies is more similar to perceiver inferences than interactant speech in standard stimulus studies, reflecting perceivers' preferential attention to their own speech (H4). The first hypothesis was supported; the remaining three were not. However, the pattern of null results may itself be informative and the exploratory analyses that followed revealed what is arguably the most consequential finding in this investigation.

The outcome measured here is cosine similarity between utterance embeddings and perceiver inference embeddings. When I describe perceiver inferences as "aligned with" or "similar to" particular utterances, this reflects measured semantic overlap, not a direct observation of perceiver attention or cognitive processing. The lens model framework organizes these similarity patterns as "cue utilization," and this framing is useful, but it rests on the assumption that semantic similarity between inferences and speech reflects something about how perceivers constructed those inferences. This assumption is plausible but not the only explanation; I return to alternative accounts and their implications in the General Discussion (Chapter V).

**Hypothesis 1: Temporal Increase in Semantic Similarity.** The first hypothesis predicted that semantic similarity between perceiver inferences and the speech stream would increase as utterances occurred closer to the moment of inference. The synthesized trajectory supported this prediction: a positive temporal trend emerged consistently across estimation methods (endpoint change and median derivative), each showing strong posterior probability of a positive trend. Both estimation methods produced credible intervals excluding zero and very strong evidence ratios in favor of a positive trend, providing consistent evidence for a temporal increase in semantic alignment across the chapter.

Importantly, the synthesized trajectory did not show the sharp spike in similarity immediately preceding the inference point that would be expected if perceivers were simply repeating the last thing they heard. Instead, the increase was gradual and distributed across the chapter, with the quintile-level decomposition revealing a modest acceleration in the final quintile rather than a discrete jump, and then only for the standard stimulus paradigm studies; the dyadic interaction data did not show the same recency effect. This distributed shape is the first piece of evidence against a simple recency heuristic and sets the stage for the subsequent hypotheses: if perceivers are not merely echoing recent speech, then with what features of the conversational stream are their inferences most aligned[26]?

**Hypotheses 2 and 3: Speaker Differences.** Hypotheses 2 and 3 addressed whether perceivers' inferences were more aligned with target speech than interactant speech, first in terms of overall similarity (H2) and then in terms of the rate of temporal change (H3). Both

[26] It is worth reiterating that cosine similarity captures semantic overlap rather than attention per se. The temporal increase may reflect conversational dynamics (speakers returning to earlier themes near topic boundaries) rather than perceiver attention patterns. I address this interpretive ambiguity more fully in the General Discussion.

hypotheses predicted a target-speech advantage: H2 predicted that target speech would show higher average similarity with perceiver inferences, and H3 predicted that both speakers would show positive temporal trends but that target speech would show a stronger increase. Neither prediction was supported by the evidence. Taken together, H2 and H3 point to a common conclusion: perceiver inferences do not systematically favor the target's speech over the interactant's. This null result, combined with the distributed temporal trajectory from H1, provides the strongest evidence against a simple recency heuristic. If perceivers were merely echoing the last thing the target said, one would expect both a pronounced recency spike and a clear target-speech advantage. Neither was observed.

The absence of a speaker difference raises the possibility that perceivers may not primarily be tracking *who* is speaking but rather *what* is being said, regardless of the source. If this is the case, the conversational context created by both speakers jointly may be more important than the contributions of either one in isolation. This interpretation aligns with work in discourse analysis showing that meaning in conversation is fundamentally co-constructed (Clark, 1996), and it suggests that inferential accuracy may depend less on monitoring a specific individual and more on comprehending the conversation as a whole.

However, the large between-study heterogeneity in interactant temporal dynamics signaled that the overall null for H3 was concealing systematic variation. Interactant speech did not behave the same way across all six studies. This heterogeneity was the primary motivation for examining paradigm as a moderator, both in the preregistered H4 test and in the subsequent exploratory analyses.

**Hypothesis 4: Interactant Mean Similarity by Paradigm.** The fourth hypothesis made a specific directional prediction that interactant speech in dyadic interaction studies would show

higher mean similarity to perceiver inferences than interactant speech in standard stimulus studies, reflecting perceivers' preferential attention to their own speech in paradigms where the perceiver is also the interactant. This prediction was not supported. The failure of H4 as a mean-level prediction points to a different conceptualization of perceiver/interactant roles in these studies. The hypothesis was predicated on the belief that perceivers in dyadic interaction studies would preferentially weight their own speech because they had produced it, but this did not account for the possibility that the *role* an interactant occupies in a conversation could matter more than the *identity* of the interactant relative to the perceiver. As the exploratory analyses reveal, paradigm type (i.e., dyadic interaction versus standard stimulus) contributes a great deal to the variability in interactant speech dynamics through temporal structure rather than overall similarity level.

### *Exploratory Analyses: Cross-Paradigm Differences*

The confirmatory hypotheses established that (a) there is a temporal increase in speech–inference similarity, (b) there is no overall speaker advantage, (c) there is no target advantage in rate of temporal change, and (d) there is no mean-level advantage for interactant speech in dyadic interaction studies. The large between-study heterogeneity in interactant temporal dynamics, however, was not explained by any of these tests. The exploratory paradigm analyses revealed where much of that heterogeneity originated, and this finding is, in my view, a key finding from Study 1.

**Temporal Dynamics Differ by Paradigm.** The temporal increase in speech–inference similarity that was strongly supported overall (H1) told two very different stories when broken down by paradigm. In standard stimulus studies, the temporal increase was decisive, with evidence ratios exceeding those of any confirmatory test. In dyadic interaction studies, the same

trend was present but weaker and more uncertain. Paradigm type was, in other words, a strong moderator of the very effect that H1 tested.

This difference was driven almost entirely by interactant speech. In standard stimulus studies, interactant speech showed the single largest temporal effect of any condition in the investigation, while target speech showed a smaller but still reliable increase. In dyadic interaction studies, neither speaker showed a temporal effect with credible intervals excluding zero. The implication is striking: the overall null for H3 (no speaker difference in rate of temporal change) was concealing a paradigm-specific pattern. Within standard stimulus studies, the speaker contrast in temporal change was large and credible, with interactant speech showing far stronger temporal dynamics than target speech.

The paradigm distinction was also clearly reflected in the GP hyperparameter estimates for interactant speech (Figure 15). In standard stimulus studies, interactant speech showed higher amplitude (larger temporal fluctuations in similarity) and shorter length-scales (more rapid change) compared to dyadic interaction studies. For target speech, GP hyperparameters did not differ meaningfully by paradigm. The $R^2\_\tau$ decomposition confirmed this selectivity: paradigm type explained the vast majority of between-study heterogeneity in interactant GP amplitude while explaining essentially none of the variance in mean similarity for either speaker or in target-specific GP hyperparameters.

This pattern tells a coherent story: paradigm moderation selectively affects the temporal structure of interactant speech rather than its overall similarity level. The GP parameters capture what average similarity cannot: in standard stimulus studies, interactant speech produces sharp, intermittent spikes of high similarity interspersed with periods of low similarity, while in dyadic interaction studies the similarity trajectory is smoother and flatter.

**Conversational Roles Versus Perceiver Identity.** The divergent trajectories across paradigm types likely follow from the different conversational roles that interactants occupy in each setting. In dyadic interaction studies, interactants and targets are social equals who share the same conversational goals (e.g., getting acquainted, as in the two Round Robin studies). In standard stimulus studies, interactants typically occupy a distinct role as interviewers or conversational facilitators who speak less frequently, and when they do speak, their contributions tend to take specific forms: asking questions that guide the conversational flow or prompting the target to elaborate on a topic.

This role distinction clarifies the GP hyperparameter estimates. Interviewer utterances are less frequent and more variable in their informational value. When an interviewer asks a question that is directly germane to the target's internal states, the resulting utterance may be highly aligned with the perceiver's subsequent inference; when the interviewer is merely prompting the target to continue speaking, the utterance may contribute less. This intermittent relevance produces the characteristic temporal signature the GP captures: high amplitude (large swings in similarity) with a short length-scale (those swings happen rapidly). The quintile decomposition is consistent with this account: the strongest interactant effects in the standard stimulus paradigm studies were concentrated in specific quintiles rather than distributed uniformly, suggesting localized elevations at particular conversational moments.

These findings suggest that the original hypotheses, which predicted that perceivers would preferentially weight their own speech in dyadic interaction studies (where the perceiver was also the interactant), did not capture the operative mechanism. The paradigm difference in interactant temporal dynamics appears to be driven not by the identity of the speaker relative to

the perceiver, but by the discursive role of the speech itself. It is the interviewer role, not the perceiver-as-interactant role, that produces the distinctive temporal pattern.

**Caveats.** Three caveats temper the paradigm finding. First, although the LOO-CV strongly favored the paradigm-moderated model for interactant GP amplitude, this improvement was not seen for most other parameters. Second, LOO-CV estimates carry high variance with only six studies, so the $R^2\_\tau$ decomposition and the LOO-CV should be read in concert rather than treated as standalone evidence. Third, the paradigm contrast may be confounded with other study-level differences (e.g., conversation length, sample characteristics, topic), and with only two standard-stimulus studies, disentangling paradigm from study-specific features is not possible. The paradigm finding should therefore be treated as strongly suggestive rather than definitive.

***Summary***

In considering Study 1 as a whole, it is worth situating the motivating concern that guided this investigation. A persistent question I have encountered when engaging with the inferential accuracy literature is whether perceivers are genuinely engaging in the complex reasoning required to infer another person's internal states, or whether above-chance accuracy scores (Ickes, 2011) may instead reflect simpler heuristics, for example, regurgitating the last thing a target said prior to the inference point. This concern is not unfounded: few individual differences reliably predict perceiver accuracy (see Chapter I), and the strongest predictor identified to date is the utilization of verbal information (e.g., Hodges & Kezer, 2021). However, prior studies that coded perceivers' use of targets' spoken words examined content within a 30–60 second window preceding each inference point without examining the specific timing of that verbal information within the stimulus. It is therefore an open question whether perceiver inferences reflect selective

engagement with the speech stream or merely default to the most temporally proximal utterances.

The counterargument that above-chance accuracy demonstrates genuine social cognition does not fully resolve this concern, as above-chance performance could arise if the most recent utterances happen to coincide with targets' reported thoughts and feelings, even if perceivers are not engaging with more diagnostic cues available earlier in the stimulus. The findings from Study 1 address this concern directly, though provide few answers, and the pattern of results across the four hypotheses, together with the exploratory paradigm analyses, suggests that the similarity data are at least somewhat inconsistent with a simple recency heuristic.

Viewed as a whole, Study 1 provides evidence that semantic similarity between perceiver inferences and the speech stream increases gradually over the course of a stimulus chapter, with a recency-like temporal effect driven primarily by interactant speech in standard stimulus studies. These similarity data are inconsistent with a simple recency heuristic: the temporal increase is distributed rather than spiked, and no overall speaker preference was detected. The most informative structural distinction was not who was speaking or temporal proximity per se, but the interaction between speaker role and paradigm type, which suggests that specific types of interactant speech (e.g., questions, topic-directing prompts) may be particularly influential when they occur in the context of an interview-style interaction.

A notable feature of these findings is that the paradigm differences in temporal dynamics would not have been detected by conventional analytic approaches. Average similarity did not differ between paradigms or speakers, and a simple linear slope would not have captured the nonlinear temporal patterns evident in the GP trajectories. The paradigm distinction emerged only when the full temporal structure of semantic similarity was modeled, revealing that

interactant speech in standard stimulus studies exhibits localized elevations in similarity that are concentrated in specific moments rather than spread uniformly across the chapter.

However, the structural analyses of Study 1 cannot directly identify which discursive features of speech account for these patterns. The substantial between-study variability in temporal dynamics, even after accounting for paradigm type, suggests that other factors, including the content and function of specific utterances, contribute to the diverse patterns observed across these six inferential accuracy studies. Study 2 addresses this gap directly. If the elevated temporal dynamics for interactant speech in standard stimulus studies reflect the interviewer's discursive role, then specific categories of interactant speech (particularly questions, which interviewers ask frequently and which can elicit self-disclosure from targets) should account for the patterns observed here. To test this, Study 2 categorizes all utterances into dialogue acts using the DAMSL coding scheme (Core & Allen, 1997) and examines whether specific dialogue act categories show differential temporal patterns in their alignment with perceiver inferences.

## III. STUDY 2: DISCURSIVE PREDICTORS OF PERCEIVER INFERENCES

Study 1 established that speech occurring later in a conversation chapter is more similar to perceiver inferences than speech occurring earlier, and that this temporal pattern depends largely on the research paradigm. However, Study 1 treated every utterance identically regardless of what kind of dialogue act it was. Not all utterances serve the same conversational function: a question invites disclosure, an informative statement shares personal content, a suggestion offers perspective, and an acknowledgment simply signals attention. If perceivers are selectively drawing on certain kinds of speech when forming their inferences, collapsing across dialogue act types could mask meaningful variation. Study 2 therefore extends the Study 1 framework by applying a new set of analyses to the same set of six studies used in Study 1 by categorizing each utterance according to its dialogue act type and asking whether certain categories of speech are more closely aligned with perceiver inferences than others.

Within the lens model framework, some dialogue acts are theoretically more likely than others to carry information that perceivers can use when forming inferences about a target's thoughts and feelings. Speech act theory (Degen, 2023; Searle, 1969) provides a foundation for these distinctions by arguing that utterances perform actions beyond conveying propositional content. The illocutionary force of an utterance (i.e., whether it asserts, requests, commits, or declares) determines the type of interpersonal work it does (Austin, 1962). Extending this framework, dialogue acts that *inform* or provide a *suggestion* involve substantive self-disclosure or novel proposals that externalize internal cognitive states, whereas dialogue acts of *acknowledgment* and *agreement* are largely reactive and carry less unique semantic content. *Question* dialogue acts occupy an intermediate position: they may not themselves express the target's internal states, but they can structure the conversational flow in ways that elicit

disclosive responses (Graesser & Person, 1994). Study 2 tests two hypotheses derived from these distinctions (cue utilization), one examining the direct alignment between specific dialogue acts and perceiver inferences (H5) and one examining the sequential influence of interactant questions on the speech that follows (H6).

**Hypothesis 5:** Specific target dialogue acts – those coded as *inform*, *suggestion*, or *question* – will show greater semantic similarity with perceiver inferences than other dialogue act categories (*acknowledgments*, *agreements*, *disagreements*, and *functional dialogue acts*). This prediction follows from the assumption that content-rich dialogue acts (i.e., those in which the speaker shares personal information, offers a novel perspective, or solicits disclosure) provide perceivers with more material from which to construct an inference about what the target is thinking or feeling.

**Hypothesis 6:** Utterances following questions from interactants will show greater semantic similarity with perceiver inferences than utterances following other interactant dialogue acts. If interactants ask questions of the target, these questions can function as conversational prompts that elicit self-disclosive responses from targets (e.g., "What was that experience like for you?"), and the speech immediately following an interactant question may be particularly aligned with what perceivers ultimately infer the targets are thinking.

As in Study 1, Stage 1 fit separate Bayesian multilevel Gaussian process models for each study, although the models are now stratified by dialogue act, and Stage 2 pooled the resulting estimates by synthesizing pointwise posterior trajectories. Because Study 1 established the analytic framework in detail, the method section below describes only the additional elements unique to Study 2: the dialogue act classification procedure, the stratification structure, and the two meta-analytic approaches used to test H5 and H6.

## Method

### *Dialogue Act Categorization*

Initial plans for this investigation sought to use the Verbal Response Modes taxonomy of dialogue acts to code the interactions in inferential accuracy stimuli (Stiles, 1992). However, upon examining this taxonomy in the specific context of inferential accuracy, the utility of Verbal Response Modes paled in comparison to the much more ubiquitous Dialogue Act Markup in Several Layers (DAMSL; Core & Allen, 1997). The reasons for this are twofold: First, the DAMSL, and particularly the DAMSL in the context of natural language processing research, was specifically developed for use with general conversational speech (e.g., Stolcke et al., 2000). In contrast, the Verbal Response Modes taxonomy of speech has largely been used in specific conversational contexts, such as supportive conversations or therapeutic interactions and thus its categories reflect the unique linguistic aspects of those spaces (e.g., Bodie et al., 2021). A more open taxonomy was therefore desirable for these heterogeneous social interactions. Second, the DAMSL is one of (if not the) most common taxonomy of dialogue acts within the area of speech coding using natural language processing (e.g., Prahallad & Mamidi, 2025). Thus, the DAMSL was also chosen to support interchangeability and generalizability to those contexts as part of the larger movement to integrate inferential accuracy research with natural language processing techniques (e.g., Kezer et al., 2025).

The DAMSL scheme used in this research is a condensed version of the set of 42 dialogue acts initially described by Jurafsky and colleagues (1997), in which related DAMSL tags were collapsed into six substantive dialogue acts based on interpersonal stance and a seventh catch-all category for tags that did not fit any substantive interpersonal function. These were built from tags with the strongest theoretical overlap to inferential accuracy and the most

frequent occurrence in everyday speech. A list of all dialogue act tags sorted into dialogue acts is shown in Table 16. All tags that didn't fit into a theoretically-meaningful dialogue act category were sorted as functional speech — this primarily included "hedges" (e.g., "I'm not sure if I'm making sense") and self-talk ("what's the word I'm looking for?"). For example, the meta-dialogue act "agreement" included the DAMSL tags "affirmative non-yes answers," "agree-accept," and "yes answers." As another example, the 42-item DAMSL has seven tags for types of questions, including "or-question" (do you want this or that?), "wh-questions" (why, who, what, etc.), and "yes-no-question" (do you know the muffin man?), which were collapsed into a single "question" dialogue act. This resulted in the following seven dialogue act categories: *acknowledgment* (7 tags), *agreement* (4 tags), *disagreement* (4 tags), *functional language* (15 tags), *inform* (3 tags), *question* (8 tags), and *suggestion* (2 tags). To ensure that the dialogue acts, which are also words that may appear in the general text of this dissertation, are clearly differentiated from their other uses, dialogue acts within the body of the document will be italicized.

To illustrate how these categories appear in the conversations used in this investigation: an *inform* act might be a target saying "I grew up in Portland and moved here for school"; a *question* act might be an interactant asking "What do you think about that?"*;* a *suggestion* act might be "You should really talk to someone in that department"*;* and an *acknowledgment* might be "Uh-huh" or "Oh, that makes sense." These categories are not mutually exclusive within a conversational turn (a single turn might contain both an *inform* and a *question*), but each individual utterance received a single dialogue act classification.

I leveraged BERT (Bidirectional Encoder Representations from Transformers; Devlin et al., 2019) models to code dialogue acts. BERT is a pretrained language model that learns

contextual representations of words by processing text in both directions simultaneously (i.e., it uses both the words that precede *and* the words that follow), enabling it to capture meaning that depends on surrounding context. This architecture has led to the creation of many derivative models that are fine-tuned for specific natural language tasks. I used four of these models in the current study trained to code DAMSL features on non-specific conversational data. These included DeepPavlov (Burtsev et al., 2018), RoBERTa (Liu et al., 2019), ConvBERT (Jiang et al., 2020), and DeBERTa (He et al., 2021). A weighted-majority voting approach was implemented such that probability estimates were generated from all four models for all categories. These were then rank order collapsed such that mutual agreement was weighted over extreme confidence from one model (see Dogan & Birant, 2019). On a holdout sample of 1,000 utterances, this approach achieved 93.44% accuracy, which is higher than single-model implementations that generally achieve between 85.8% and 91.6% accuracy (Żelasko et al., 2021).

**Table 16:**

*DAMSL Dialogue Act Tags Sorted into Discursive Categories from Jurafsky et al. (1997)*

| Study Category | DAMSL Code | DAMSL Tag Name | Example |
|---|---|---|---|
| **Acknowledgment** | b | Acknowledge (Backchannel) | *Uh-huh.* |
| | bk | Response Acknowledgement | *Oh, okay.* |
| | bh | Backchannel in question form | *Is that right?* |
| | b^m | Repeat-phrase | *Oh, fajitas* |
| | ba | Appreciation | *I can imagine.* |
| | bd | Downplayer | *That's all right.* |
| | bf | Summarize/reformulate | *Oh, you mean you switched schools.* |
| **Agreement** | aa | Agree/Accept | *That's exactly it.* |
| | aap/am | Maybe/Accept-part | *Something like that* |
| | ny | Yes answers | *Yes.* |
| | na | Affirmative non-yes answers | *It is.* |
| **Disagreement** | ar | Reject | *Well, no* |
| | nn | No answers | *No.* |
| | ng | Negative non-no answers | *Uh, not a whole lot.* |
| | arp/nd | Dispreferred answers | *Well, not so much that.* |
| **Functional** | % | Abandoned or Turn-Exit | *So, -* |
| | % | Uninterpretable | *But, uh, yeah* |
| | x | Non-verbal | *[Laughter]* |
| | fc | Conventional-closing | *Well, it's been nice talking to you.* |
| | fp | Conventional-opening | *How are you?* |
| | h | Hedge | *I don't know if I'm making any sense.* |
| | ^h | Hold before answer/agreement | *I'm drawing a blank.* |
| | br | Signal-non-understanding | *Excuse me?* |
| | t1 | Self-talk | *What's the word I'm looking for* |
| | t3 | 3rd-party-talk | *My goodness, Diane, get down.* |
| | o/fo/bc/by/fw | Other | *Well give me a break, you know.* |
| | ft | Thanking | *Hey thanks a lot* |

| Study Category | DAMSL Code | DAMSL Tag Name | Example |
|---|---|---|---|
| | fa | Apology | *I'm sorry.* |
| | ^2 | Collaborative Completion | *Who aren't contributing.* |
| | no | Other answers | *I don't know* |
| **Inform** | sd | Statement-non-opinion | *Me, I'm in the legal department.* |
| | sv | Statement-opinion | *I think it's great* |
| | ^q | Quotation | *You can't be pregnant and have cats* |
| **Question** | qy | Yes-No-Question | *Do you have any special training?* |
| | qw | Wh-Question | *Well, how old are you?* |
| | qo | Open-Question | *How about you?* |
| | qh | Rhetorical-Questions | *Who would steal a newspaper?* |
| | qy^d | Declarative Yes-No-Question | *So you can afford a house?* |
| | qw^d | Declarative Wh-Question | *You are what kind of buff?* |
| | qrr | Or-Clause | *or is it more of a company?* |
| | ^g | Tag-Question | *Right?* |
| **Suggestion** | ad | Action-directive | *Why don't you go first* |
| | oo/cc/co | Offers, Options, Commits | *I'll have to check that out* |

### *Analytic Approach*

The analytic pipeline for Study 2 followed the same two-stage individual participant data meta-analysis framework described in the Study 1 Method section. Stage 1 fit separate Bayesian multilevel Gaussian process models (Equation 3) for each dialogue act within each study. Stage 2 followed the same meta-analytic pooling sequence: Stage 2A univariate random-effects meta-analyses, Stage 2B multivariate trajectory synthesis via spline-based dimension reduction, and Stage 2C paradigm-moderated meta-analyses. All model specifications, prior distributions, and sensitivity analysis procedures were identical to those described for Study 1. Utterances were

temporally indexed using the same proportional time scale described in the Study 1 Method, with each utterance's timestamp normalized to the [0, 1] interval within its chapter.

The key structural difference from Study 1 is the addition of dialogue act as a stratifying variable. I thus conducted Stage 2A meta-analyses using two complementary approaches: Approach A pooled all estimates in a single multilevel model with dialogue act as a fixed-effect moderator (yielding overall heterogeneity estimates and enabling formal tests of the H5 focal contrasts: *inform*, *suggestion*, *question* vs. baseline acts), while Approach B fit separate meta-analyses per dialogue act (yielding dialogue-act-specific pooled estimates).

**Results**

***Stage 1: Individual Bayesian Multilevel Models***

I used the same model equation (Equation 3) to estimate the temporal distribution of speech-inference similarity (the cosine similarity between utterance embeddings and perceiver inference embeddings; higher values indicate that what a perceiver wrote about what the target was thinking more closely matched the words actually spoken) for each dialogue act across interactant and target speech within each of the six studies. This resulted in 35 individual models for Study 2. Although 42 models were specified (7 dialogue acts by 6 studies), seven dialogue act-by-study combinations lacked sufficient data for the multilevel model to estimate at least one level (i.e., in two studies, interactants did not disagree enough to estimate the level of the model that involved the *disagreement* dialogue act). Individual model fit and model estimates are shown in Appendix A. Model fit statistics summarized by dialogue act are shown in Table 17.

**Table 17:**

*Study 2: Individual Bayesian Multilevel Model Convergence Statistics*

| Dialogue Act | N Studies | Max R-hat | Total Observations |
|---|---|---|---|
| Acknowledgment | 5 | 1.01 | 788 |
| Agreement | 6 | 1.00 | 35,125 |
| Disagreement | 4 | 1.00 | 13,065 |
| Functional | 4 | 1.00 | 4,743 |
| Inform | 6 | 1.01 | 88,402 |
| Question | 6 | 1.00 | 28,641 |
| Suggestion | 4 | 1.00 | 801 |

### *Stage 2A: Univariate Meta-Analyses*

I followed the same meta-analytic approach in Study 2 as Study 1 by first specifying univariate meta-analytic models for the three parameters of interest: fixed effect of speaker type and the two Gaussian hyperparameters. These models follow the same likelihood functions as those in Stage 2A of Study 1 (Equations 12 & 13). Results for the meta-analysis model estimating the average speech-inference similarity for each dialogue act (by speaker) found no substantial difference as a function of speaker type for any dialogue act (Table 18). In other words, whether an *inform* or *question* dialogue act was spoken by the target or the interactant did not systematically change how closely the words uttered semantically matched what the perceiver ultimately wrote as their inference about the target's thoughts.

**Table 18:**

*Study 2: Fixed Effects of Dialogue Act by Speaker Type with 95% CIs*

| Dialogue Act | μ Target [95% CI] | μ Interactant [95% CI] | μ Diff (T−P) [95% CI] | P(μ T > P) | Max Rhat | Min ESS |
|---|---|---|---|---|---|---|
| Acknowledgment | 0.098 [0.031, 0.169] | 0.112 [0.048, 0.152] | -0.014 [−0.092, 0.083] | 0.317 | 1.00 | 1,396.41 |
| Agreement | 0.119 [0.104, 0.127] | 0.121 [0.103, 0.138] | -0.002 [−0.024, 0.017] | 0.389 | 1.00 | 2,188.95 |
| Disagreement | 0.125 [0.001, 0.188] | 0.132 [0.042, 0.331] | -0.007 [−0.186, 0.108] | 0.475 | 1.00 | 1,217.77 |
| Functional | 0.136 [0.063, 0.209] | 0.137 [0.075, 0.195] | -0.001 [−0.097, 0.097] | 0.465 | 1.01 | 1,266.45 |
| Inform | 0.130 [0.097, 0.160] | 0.133 [0.087, 0.176] | -0.003 [−0.057, 0.051] | 0.437 | 1.01 | 839.27 |
| Question | 0.141 [0.072, 0.203] | 0.138 [0.074, 0.198] | 0.003 [−0.087, 0.091] | 0.521 | 1.00 | 1,619.51 |
| Suggestion | 0.159 [0.047, 0.265] | 0.151 [0.074, 0.219] | 0.009 [−0.131, 0.139] | 0.580 | 1.01 | 1,201.35 |

*μ = pooled mean cosine similarity. Diff = Target − Interactant. P(T > P) = posterior probability that target exceeds interactant.*

Table 19 shows the results from the meta-analysis of Gaussian process amplitudes (how much speech-inference similarity fluctuates over the course of a chapter) for each speaker type and dialogue act. No substantial difference between speakers was found for any dialogue act in the univariate amplitude meta-analysis.

**Table 19:**

*Study 2: Dialogue Act Gaussian Process Amplitude Estimates with 95% CIs*

| Dialogue Act | Amplitude Target [95% CI] | Amplitude Interactant [95% CI] | Amplitude Ratio (T/P) [95% CI] | P(T > P) |
|---|---|---|---|---|
| Acknowledgment | 0.674 [0.187, 1.720] | 0.702 [0.181, 1.756] | 0.946 [0.205, 4.845] | 0.473 |
| Agreement | 0.655 [0.145, 1.718] | 0.696 [0.191, 1.762] | 0.930 [0.176, 4.107] | 0.464 |
| Disagreement | 0.814 [0.258, 1.879] | 0.813 [0.257, 1.960] | 1.001 [0.234, 4.204] | 0.500 |
| Functional | 0.815 [0.253, 1.930] | 0.819 [0.257, 1.965] | 0.985 [0.227, 4.411] | 0.489 |
| Inform | 0.648 [0.194, 1.632] | 0.694 [0.229, 1.634] | 0.913 [0.214, 3.856] | 0.450 |
| Question | 0.720 [0.229, 1.718] | 0.684 [0.222, 1.718] | 1.057 [0.252, 4.454] | 0.533 |
| Suggestion | 0.747 [0.182, 1.912] | 0.689 [0.145, 1.692] | 1.078 [0.213, 6.006] | 0.534 |

*Amplitude (sdgp) = log transformed GP amplitude (magnitude of temporal variation); Ratio > 1 indicates target trajectories are smoother. Back-transformed from log scale.*

Table 20 shows the length-scale differences (how smoothly the similarity pattern changes over the chapter) for each dialogue act and, as in the previous models, no substantial differences were observed across speaker type. Across the three univariate meta-analyses, no clear speaker differences were seen for any dialogue act. However, there was a large amount of variability in the estimates, suggesting that there may be differences within dialogue acts that are not captured by the univariate meta-analyses.

**Table 20:**

*Study 2: Dialogue Act Gaussian Process Length-Scale Estimates with 95% CIs*

| Dialogue Act | Length-scale Target [95% CI] | Length-scale Interactant [95% CI] | Length-scale Ratio (T/P) [95% CI] | P(T > P) |
|---|---|---|---|---|
| Acknowledgment | 0.680 [0.285, 1.419] | 0.749 [0.286, 1.664] | 0.936 [0.279, 3.021] | 0.459 |
| Agreement | 0.768 [0.223, 1.858] | 0.708 [0.202, 1.720] | 1.099 [0.224, 4.962] | 0.542 |
| Disagreement | 0.950 [0.339, 2.187] | 0.781 [0.289, 1.769] | 1.234 [0.331, 4.412] | 0.605 |
| Functional | 1.003 [0.352, 2.365] | 1.013 [0.366, 2.244] | 0.996 [0.266, 3.678] | 0.497 |
| Inform | 0.690 [0.219, 1.621] | 0.707 [0.228, 1.673] | 0.972 [0.232, 4.059] | 0.488 |
| Question | 0.715 [0.227, 1.756] | 0.761 [0.234, 1.875] | 0.941 [0.221, 4.231] | 0.464 |
| Suggestion | 0.926 [0.327, 2.164] | 0.761 [0.282, 1.674] | 1.182 [0.324, 4.516] | 0.606 |

*Length-scale (lscale) GP log transformed length-scale (trajectory smoothness; larger = smoother). Ratio > 1 indicates target trajectories are smoother. Back-transformed from log scale.*

**Between-Study Heterogeneity.** The dialogue-act-stratified meta-analyses present a more layered picture of between-study heterogeneity than the structural models in Study 1 because

each dialogue act is meta-analyzed jointly via a multilevel moderator model (Approach A) and in separate, individual models (Approach B). I examined τ under both approaches. The Approach A distributions, which estimate a single between-study standard deviation across all dialogue acts within each speaker type, are shown as ridge plots in Figure 17. The Approach B distributions (which allowed τ to vary by dialogue act) are shown as interval plots in Figure 18.

Under the Approach A moderator model, fixed effect heterogeneity was modest for both speaker types, broadly consistent with what was observed for the overall structural models in Study 1. The mean level of semantic similarity between specific dialogue act categories and perceiver inferences was relatively stable across the six studies—even though the studies differed substantially in paradigm type, conversational content, and sample composition.

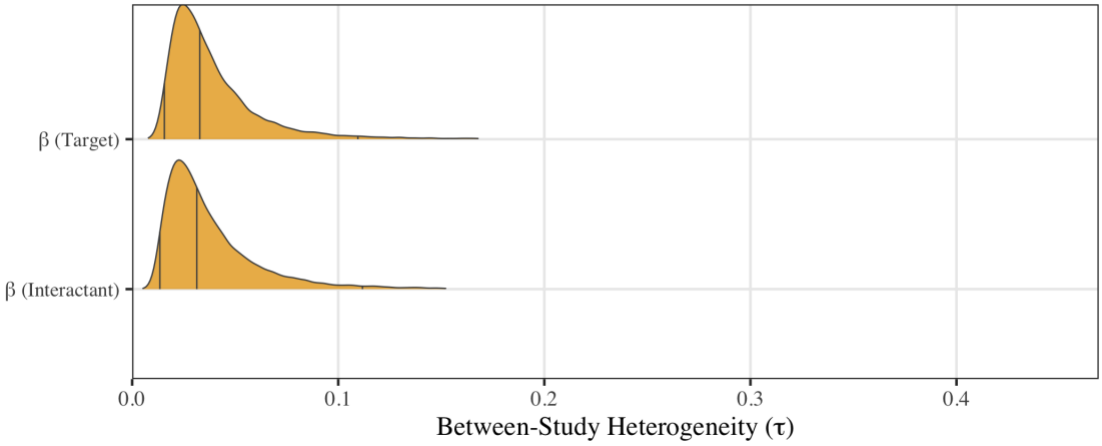
The per-dialogue-act results from Approach B (Figure 18) add useful granularity. Fixed effect heterogeneity was generally small and consistent across dialogue acts, with two exceptions: *acknowledgment* and *suggestion* (the two lowest-frequency categories) showed wider τ credible intervals, likely reflecting greater estimation uncertainty rather than genuinely elevated between-study variability. More interesting was the pattern in GP amplitude heterogeneity, where *agreement* and *question* dialogue acts showed notably larger τ values. This suggests that temporal fluctuations in these high-frequency dialogue acts were particularly sensitive to study-specific features. The paradigm moderation analyses help explain why: *agreement* and *question* amplitudes differed substantially between standard stimulus and dyadic interaction studies, so pooling across paradigms without accounting for this distinction produces inflated τ.

The dialogue-act-stratified Approach B models allow a more granular decomposition of between-study heterogeneity by paradigm than is possible with the overall structural models. For each dialogue act × speaker × parameter combination, I computed the posterior distribution of

$R^2_\tau$ by pairing the unconditional per-DA meta-analytic fits with their paradigm-moderated counterparts.

**Figure 17:**

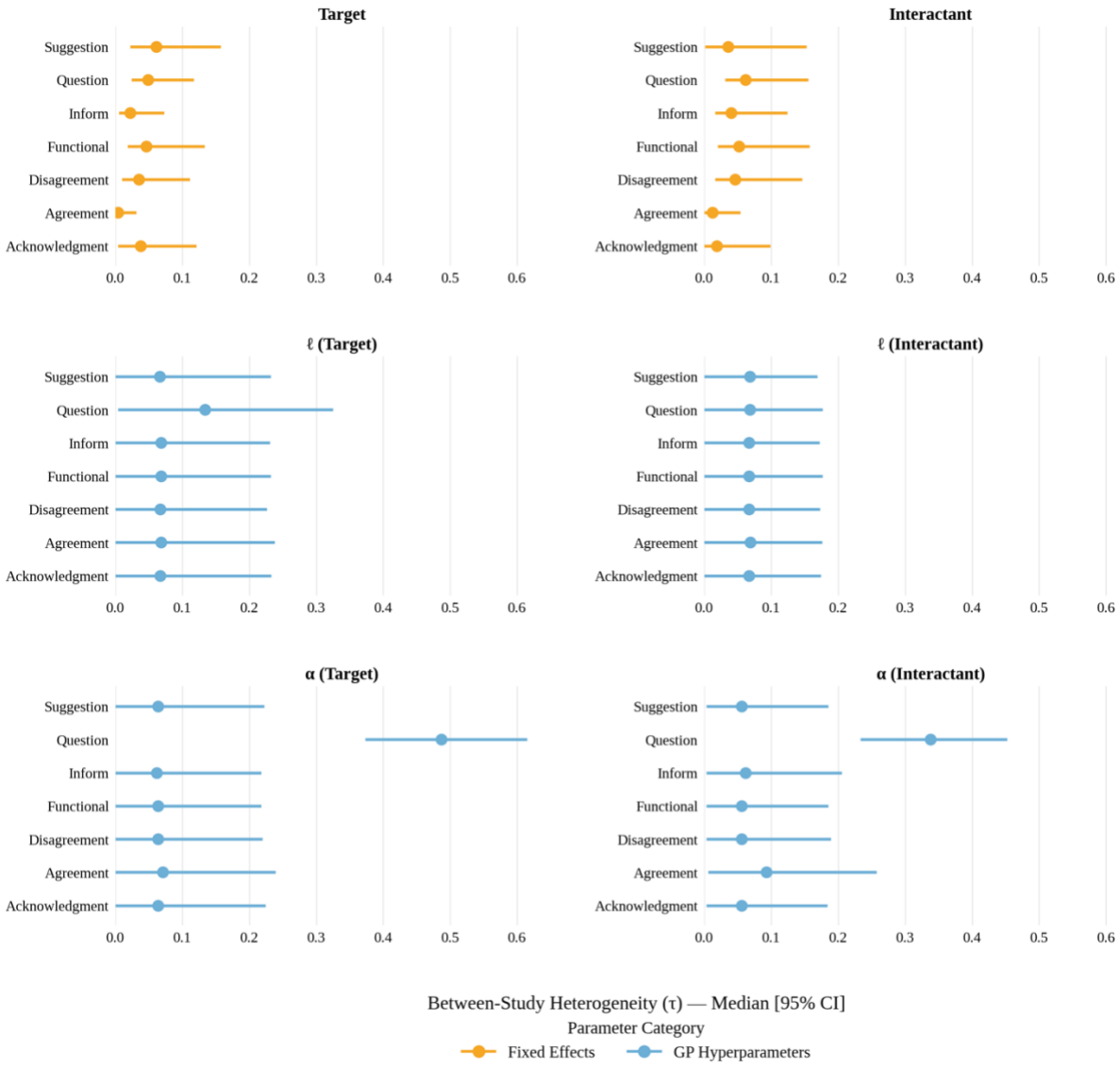
*Study 2: H5 Between-Study Heterogeneity (τ): Approach A Posterior Distributions*




The per-dialogue-act decomposition revealed that the proportion of heterogeneity attributable to paradigm varied considerably across dialogue acts. For the fixed effects, $R^2_\tau$ was generally modest and broadly similar across dialogue acts, consistent with the finding that mean-level semantic similarity was relatively stable regardless of paradigm. The GP amplitude decomposition was more revealing: agreement and question acts showed the largest $R^2_\tau$ values, indicating that the paradigm-related heterogeneity in temporal dynamics was concentrated in these high-frequency, interactionally consequential dialogue acts.

**Figure 18:**

*Study 2: H5 Between-Study Heterogeneity (τ) by Dialogue Act: Approach B Point Estimates and 95% Cis*




The LOO-CV comparisons for the individual models were generally inconclusive, as expected given the small number of studies per dialogue act (4-6). Where the $R^2\_\tau$ decomposition identified substantial paradigm effects (e.g., *agreement* and *question* amplitude),

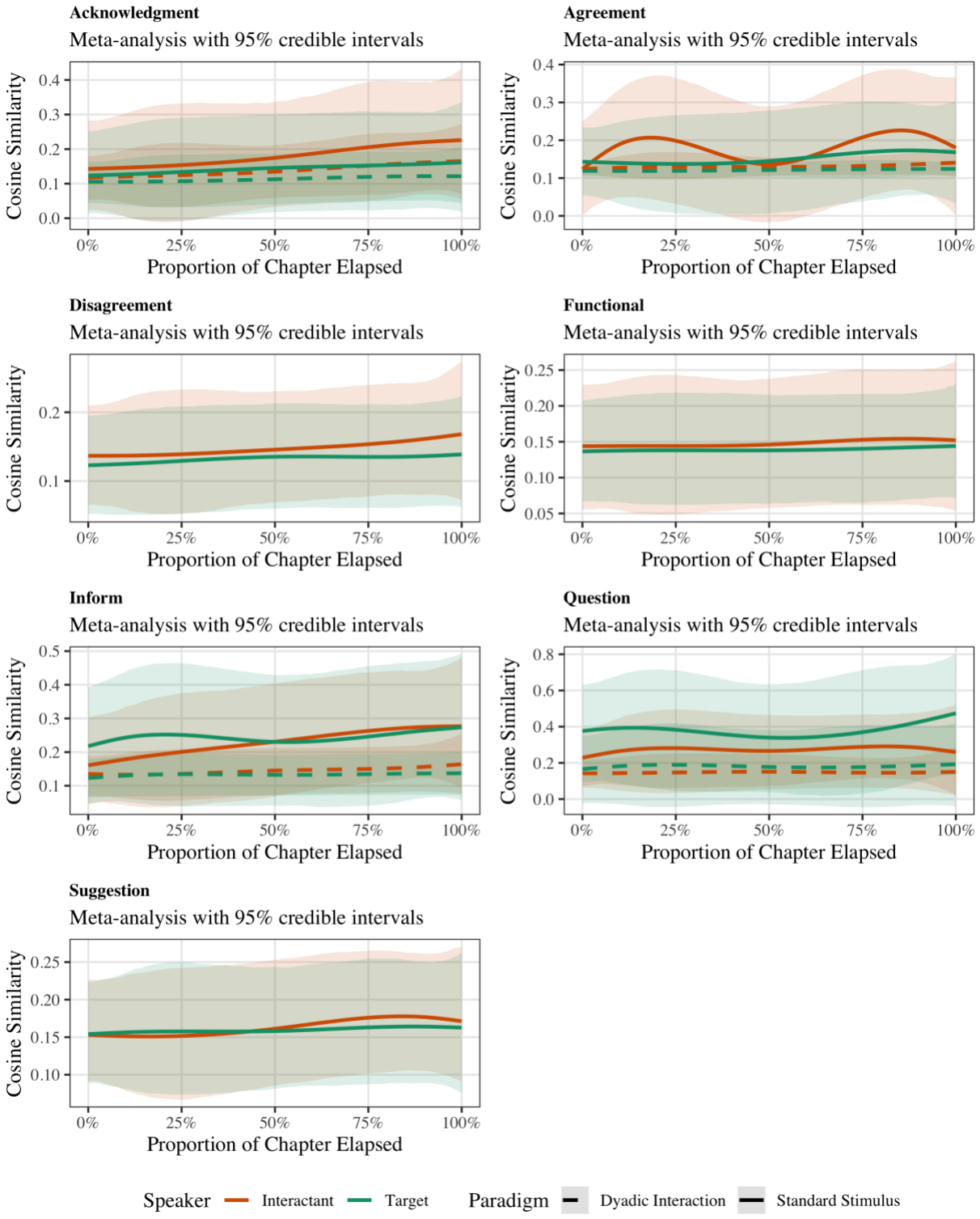
the LOO-CV results tended weakly in the same direction, but the standard errors were too large to draw confident conclusions from the predictive comparison alone.

### *Stage 2B: Multivariate Synthesis of Pointwise Posterior Trajectories*

I used the same process for creating the pooled pointwise posterior trajectories in Study 2 as Study 1 by first creating an approximation of each individual study's posterior trajectory for both targets and interactants. I then used these functions to pool the estimates for all draws from the posterior, creating a combined posterior trajectory distribution for each speaker type and dialogue act. Figure 19 shows the estimated cosine similarity for each dialogue act across the chapter with 95% credible intervals; as can be seen, the standard stimulus studies lacked sufficient observations of *disagreement*, *functional*, and *suggestion* dialogue acts for the models to meet convergence criteria and were thus excluded from paradigm-specific moderation analyses. For *acknowledgment*, *disagreement*, and *functional* dialogue acts, there is very little difference between speakers at any time point in the video chapter. However, *inform* and *question dialogue acts* show notable trajectory shapes across the chapter that are examined in more detail below.

**Figure 19:**

*Study 2: Synthesized GP Trajectories by Dialogue Act*



### *Stage 2C: Moderated Meta-Analyses*

The picture becomes even more interesting when the data are further divided by their paradigm: dyadic interaction or standard stimulus. Estimates and contrasts for dialogue act by speaker type by paradigm are shown below for the four dialogue acts with sufficient standard stimulus study observations for the analyses (*acknowledgment, agreement, inform, question*): Table 21 shows the contrasted fixed effects of speaker type and Tables 22 and 23 shows the contrasted Gaussian process hyperparameters. Although there is suggestive evidence that target *agreement* dialogue acts are more semantically similar to perceiver inferences in the dyadic interaction paradigm than in the standard stimulus paradigm, the clearer differences are seen in the Gaussian process hyperparameters.

**Table 21:**

*Study 2: Paradigm × Speaker Differences in Dialogue Act Mean Estimates with 95% CIs*

| Dialogue Act | Speaker | μ SS [95% CI] | μ DI [95% CI] | μ Δ(DI−SS) [95% CI] | P(μ DI > SS) |
|---|---|---|---|---|---|
| Acknowledgment | Target | 0.113 [0.030, 0.200] | 0.106 [0.083, 0.129] | -0.007 [−0.101, 0.080] | 0.450 |
| | Interactant | 0.088 [0.003, 0.169] | 0.117 [0.095, 0.140] | 0.029 [−0.056, 0.115] | 0.750 |
| Agreement | Target | 0.074 [0.017, 0.134] | 0.120 [0.115, 0.124] | 0.045 [−0.014, 0.104] | 0.934 |
| | Interactant | 0.111 [0.023, 0.200] | 0.120 [0.116, 0.124] | 0.009 [−0.080, 0.098] | 0.572 |
| Inform | Target | 0.135 [0.091, 0.176] | 0.117 [0.112, 0.122] | -0.018 [−0.059, 0.026] | 0.196 |
| | Interactant | 0.195 [0.082, 0.301] | 0.127 [0.121, 0.132] | -0.068 [−0.174, 0.044] | 0.116 |
| Question | Target | 0.135 [0.038, 0.232] | 0.118 [0.112, 0.124] | -0.017 [−0.114, 0.081] | 0.358 |
| | Interactant | 0.148 [0.074, 0.222] | 0.114 [0.109, 0.120] | -0.034 [−0.108, 0.040] | 0.201 |

*SS = standard stimulus studies. DI = dyadic interaction studies. Δ = difference (DI − SS). P(DI > SS) = posterior probability that DI exceeds SS.*

When split by paradigm type, *agreement*, *inform*, and *question* dialogue acts all show substantial inter-paradigm differences. For perceivers in the standard stimulus paradigm, the amplitude of these three dialogue acts is much higher than in the dyadic interaction paradigm (meaning similarity fluctuates more over the chapter), while the length-scale is substantially smaller (meaning those fluctuations are more rapid and localized). In practical terms, when perceivers watch a standard stimulus interview, the alignment between specific inform, agreement, and question utterances and perceivers' inferences varies sharply from one part of the chapter to the next. In dyadic interaction paradigm studies, the alignment is comparatively flat. This suggests that the interview structure of standard stimulus paradigm studies creates windows in the conversation where certain dialogue acts become especially informative to perceivers, rather than dialogue acts being uniformly weighted throughout.

**Table 22:**

*Study 2: Paradigm × Speaker Differences in Dialogue Act Amplitude Estimates with 95% CIs*

| Dialogue Act | Speaker | Standard Stimulus [95% CI] | Dyadic Interaction [95% CI] | Dyadic Interaction − Standard Stimulus [95% CI] | P(Dyadic Interaction > Standard Stimulus) |
|---|---|---|---|---|---|
| Acknowledgment | Interactant | 0.038 [0.009, 0.163] | 0.028 [0.011, 0.072] | -0.308 [−2.024, 1.408] | 0.368 |
| | Target | 0.031 [0.008, 0.115] | 0.027 [0.011, 0.066] | -0.131 [−1.814, 1.442] | 0.445 |
| Agreement | Interactant | 0.129 [0.080, 0.205] | 0.015 [0.005, 0.044] | -2.133 [−3.252, −0.947] | < 0.001 |
| | Target | 0.178 [0.125, 0.250] | 0.015 [0.005, 0.045] | -2.491 [−3.621, −1.282] | < 0.001 |
| Inform | Interactant | 0.107 [0.069, 0.165] | 0.024 [0.010, 0.059] | -1.509 [−2.416, −0.527] | 0.002 |
| | Target | 0.506 [0.383, 0.666] | 0.018 [0.009, 0.039] | -3.339 [−4.117, −2.505] | < 0.001 |
| Question | Interactant | 0.960 [0.725, 1.267] | 0.024 [0.011, 0.050] | -3.709 [−4.510, −2.862] | < 0.001 |
| | Target | 0.383 [0.281, 0.524] | 0.018 [0.007, 0.045] | -3.075 [−3.996, −2.137] | < 0.001 |

*Estimates back-transformed from log scale. Difference is on log scale.*

**Table 23:**

*Study 2: Paradigm × Speaker Differences in Dialogue Act Length-Scale Estimates with 95% CIs*

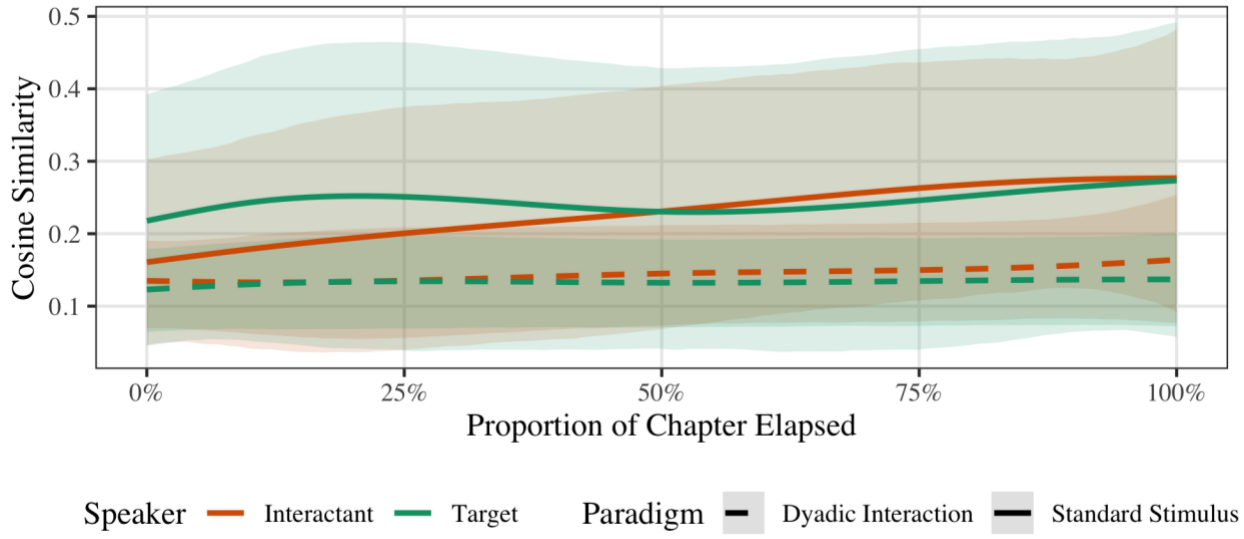
| Dialogue Act | Speaker | Standard Stimulus [95% CI] | Dyadic Interaction [95% CI] | Dyadic Interaction − Standard Stimulus [95% CI] | P(Dyadic Interaction > Standard Stimulus) |
|---|---|---|---|---|---|
| Acknowledgment | Interactant | 0.384 [0.118, 1.214] | 0.200 [0.045, 0.820] | -0.653 [−2.549, 1.275] | 0.244 |
| | Target | 0.320 [0.062, 1.549] | 0.248 [0.052, 1.087] | -0.258 [−2.512, 2.021] | 0.400 |
| Agreement | Interactant | 0.051 [0.024, 0.100] | 0.354 [0.060, 2.307] | 1.941 [0.087, 3.961] | 0.981 |
| | Target | 0.076 [0.040, 0.142] | 0.325 [0.051, 2.181] | 1.454 [−0.543, 3.502] | 0.937 |
| Inform | Interactant | 0.046 [0.023, 0.087] | 0.278 [0.060, 1.516] | 1.799 [0.095, 3.575] | 0.982 |
| | Target | 0.061 [0.038, 0.097] | 0.214 [0.059, 0.795] | 1.257 [−0.160, 2.683] | 0.957 |
| Question | Interactant | 0.059 [0.040, 0.084] | 0.269 [0.047, 1.620] | 1.514 [−0.316, 3.364] | 0.954 |
| | Target | 0.048 [0.031, 0.078] | 0.327 [0.054, 2.079] | 1.915 [0.076, 3.806] | 0.980 |

*Estimates back-transformed from log scale. Difference is on log scale.*

The pattern implied by this set of amplitude and length-scale hyperparameters is most clearly visible in the pointwise posterior predictions in Figures 20 and 21. For *inform* dialogue acts (Figure 20), the standard stimulus paradigm reveals a convergence pattern between speakers across the video chapter. The target *inform* trajectory begins at an elevated level of similarity (approximately 0.23) and remains relatively stable or rises slightly across the chapter, while the interactant *inform* trajectory begins notably lower (approximately 0.14) and increases steadily, approaching the target trajectory by the chapter's end. The separation between speakers is thus widest in the opening portion of the chapter and narrows progressively — though it is important to note that the 95% credible intervals for both trajectories are wide and overlapping, indicating substantial uncertainty in the exact trajectory shapes at any given time point. In the dyadic interaction paradigm, by contrast, both speaker trajectories remain flat and essentially overlapping. To the extent that these posterior means are taken at face value, the pattern suggests that in standard stimulus studies, target inform utterances maintain a consistently moderate level of alignment with perceiver inferences throughout the chapter, while *inform* interactant utterances, which were initially less aligned, become progressively more similar to perceiver inferences closer to the inference point.

***Figure 20:***

*Study 2: Inform Dialogue Acts Pointwise Posterior Predictions by Paradigm*

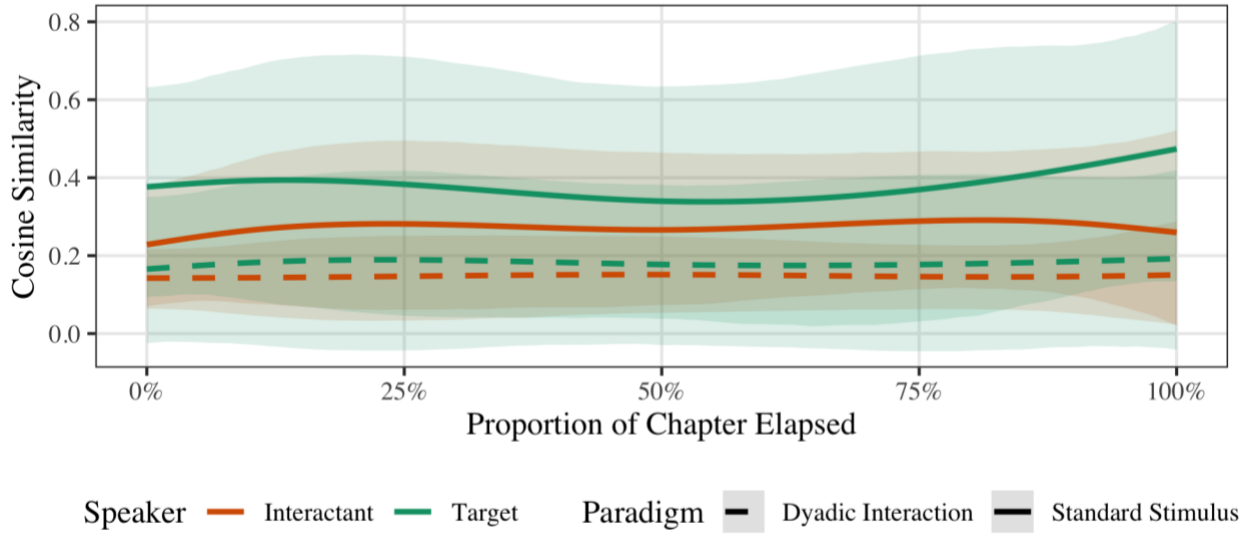



For *question* dialogue acts (Figure 21), the standard stimulus paradigm suggests that there may be a modest increase in similarity later in the chapter: the posterior mean trajectories suggest that semantic similarity between *question* utterances and perceiver inferences increases across the video chapter for both speakers. Both target and interactant *question* trajectories begin at moderate levels of similarity (approximately 0.20–0.25) and rise across the chapter, with target *questions* reaching approximately 0.35 and interactant *questions* reaching approximately 0.22–0.25 by the chapter's end. The two trajectories rise in a roughly parallel fashion, with the target trajectory consistently somewhat higher. However, the 95% credible intervals surrounding both trajectories are notably wide and span from near zero to above 0.60 at some time points, reflecting the considerable between-study heterogeneity. These wide intervals suggest that this apparent end-of-chapter rise, while consistent with the direction implied by the GP hyperparameters, should be interpreted very cautiously. In contrast, as with *inform* acts, the dyadic interaction paradigm trajectories are flat and undifferentiated. To the extent that the

posterior means capture a genuine temporal pattern, the rising similarity for *question* dialogue acts in the standard stimulus paradigm studies suggests that questions occurring later in an interview may be more closely aligned with what perceivers ultimately infer than those earlier in the chapter.

**Figure 21:**

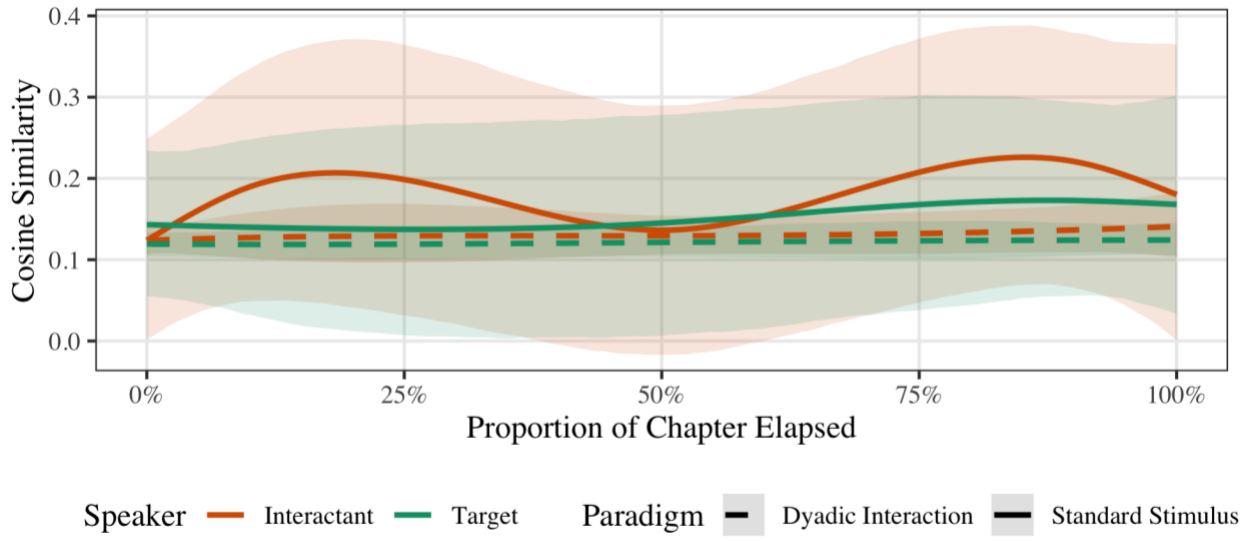
*Study 2: Question Dialogue Acts Pointwise Posterior Predictions by Paradigm*




*Agreement* dialogue acts (Figure 22) present a less differentiated picture. The interactant agreement trajectory shows a modest elevation in the middle portion of the chapter, though the credible intervals overlap throughout and the separation between speakers is neither pronounced nor temporally localized in the way that inform and question acts are in the standard stimulus paradigm. This suggests that while *agreement* utterances may contribute to the general flow of conversational alignment, they do not show the kind of speaker-specific or temporally structured signatures visible in the other content-bearing dialogue acts.

**Figure 22:**

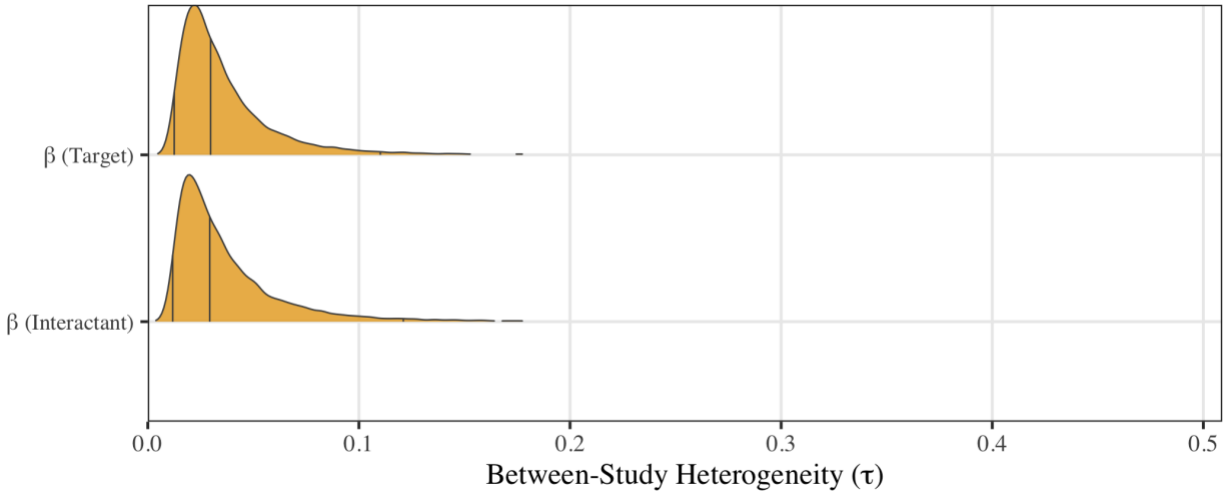
*Study 2: Agreement Dialogue Acts Pointwise Posterior Predictions by Paradigm*




**Heterogeneity Decomposition (H6).** The between-study heterogeneity patterns for H6 largely mirrored those for H5. Fixed effect τ values (Approach A) were again modest (Figure 23), indicating that the association between prior-turn dialogue acts and semantic similarity was stable across studies.

***Figure 23:***

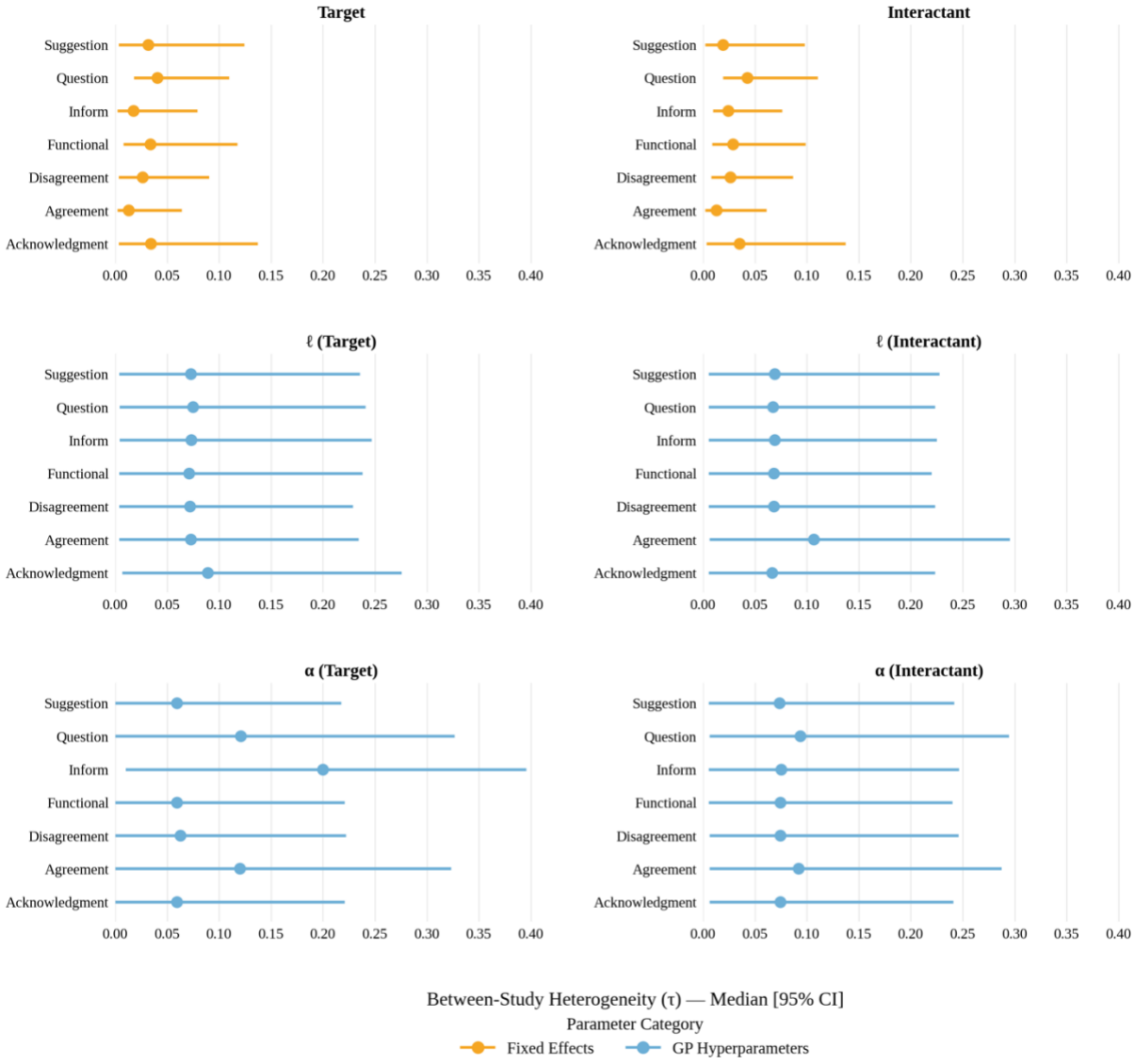
*H6 Between-Study Heterogeneity (τ): Approach A Posterior Distributions*




The caterpillar plots (Approach B; Figure 24) showed generally comparable heterogeneity across dialogue act categories, though *question* dialogue acts showed somewhat elevated τ for GP amplitude. This is worth noting because H6 specifically examines the sequential influence of questions, and the elevated heterogeneity suggests that the temporal consequences of asking a question may depend on study-specific features of the conversational context.

**Figure 24:**

*Study 2: H6 Between-Study Heterogeneity (τ) by Dialogue Act: Approach B Point Estimates and 95% CIs*




## Discussion

Study 2 extended the structural analysis of Study 1 by asking whether the content of specific types of speech are more similar than others to perceivers' inferences. The answer seems to be conditional: not all dialogue acts are created equal, but the distinctions that matter may

depend in part on when the utterance occurs and under what paradigm. *Inform* and *question* dialogue acts showed evidence of distinct temporal profiles for targets versus interactants, with the clearest speaker-type differences emerging in the boundary regions of the chapter (the first and last portions) and primarily within the standard stimulus paradigm. In this discussion, I review the evidence for two specific hypotheses addressed in Study 2 and then turn to the broader implications of these findings for understanding how perceivers may navigate the verbal landscape of a conversation.

As in Study 1, the outcome measured was cosine similarity between utterance embeddings and perceiver inference embeddings. Although interpreting these differences as reflecting perceiver attention to or utilization of specific dialogue acts is reasonable within the lens model framework, it is worth reiterating that perceivers' inferences and target and interactant speech may converge semantically for reasons other than attentional processing. Semantic similarity is not a causal measure.

### *Hypothesis 5: Target Inform, Suggestion, and Question Dialogue Acts Will Predict Higher Speech-Inference Similarity*

Hypothesis 5 predicted that if perceivers selectively attend to content-rich dialogue acts, then *inform*, *suggestion*, and *question* utterances by targets should show higher cosine similarity with perceiver inferences than other dialogue act categories. This prediction was only partially supported. The expected global elevation in semantic similarity for *inform* and *question* target utterances was not observed as a mean-level effect; however, the trajectory synthesis revealed suggestive evidence for temporally localized patterns that a mean-level analysis would miss. These patterns were confined to the standard stimulus paradigm.

The *inform* and *question* findings present a nuanced picture when set against the Study 1 temporal increase. For *inform* acts, it is not that early informative speech is uniquely aligned with perceiver inferences in the sense of a declining trajectory; rather, target *inform* dialogue acts maintain a stable level of alignment throughout while interactant *inform* dialogue acts gradually rise to converge with it. For *question* acts, the pattern is more straightforwardly consistent with the gradual temporal increase: *questions* later in the standard stimulus chapter are more aligned with perceiver inferences than *questions* earlier in the chapter, paralleling the overall temporal pattern from Study 1 at the level of a specific dialogue act category. Together, these patterns suggest that the interview structure of the standard stimulus paradigm creates distinct temporal signatures for different dialogue acts, rather than a single temporal pattern that applies uniformly.

***Hypothesis 6: Interactant Question Dialogue Acts Predict Higher Speech-Inference Similarity***

Hypothesis 6 predicted that *question* dialogue acts from the interactant (e.g., *"What was that experience like for you?"*) would elicit target disclosures that align with perceiver inferences, producing higher speech-inference similarity for utterances following interactant *question* dialogue acts. This hypothesis was not supported. No unique effect of interactant *question* dialogue acts was found in any model, paradigm, or portion of the video chapter. The average similarity for interactant *question* dialogue acts was indistinguishable from that of other dialogue acts, and the trajectory synthesis showed no time-dependent divergence between speakers for *question* dialogue acts when interactant speech was the focus.

The null result is surprising given the intuitive appeal of the hypothesis and the robust finding that target *question* dialogue acts do show temporal modulation, in that semantic similarity seems highest for *question* utterances occurring later in the chapter. One possible explanation is that interactant *question* dialogue acts vary enormously in content and thus also

what they elicit. An open-ended *question* dialogue act (*"What made you decide to change majors?"*) may prompt genuine self-disclosure, while an emphatic *question* dialogue act (*"Really?"* or *"How about you?"*) may not. Because my application of the DAMSL coding scheme collapsed all question types into a single category, these functionally distinct dialogue acts were averaged together, potentially masking a real effect for substantive *question* dialogue acts.

A second possibility is that what matters for perceiver inferences is not the *question* itself but the response it generates. If a perceiver's inference aligns with the target's subsequent disclosure rather than with the question that prompted it, the *question*-similarity association may be null even when a *question* dialogue act played a causal role in the conversation. For example, if a participant was asked the name of their childhood pet, they may explain that their pet was named after a favorite television character and then proceed to describe why they were drawn to that character or television program. Although the *question* dialogue act was directly involved in eliciting the response, the content would not necessarily be semantically similar. Future research could address this by coding *question* dialogue act subtypes (open vs. closed, substantive vs. emphatic) and examining the similarity of the target's response to the perceiver's inference conditional on the preceding *question* type.

***Dyadic Interaction and Standard Stimulus Paradigm Asymmetry***

The restriction of these dialogue act effects to the standard stimulus paradigm warrants additional comment. In standard stimulus studies, the interactant is typically more of an interviewer following a structured or semi-structured protocol and the target is more like an interviewee responding to the interactant's questions. This asymmetry in conversational roles may create a more predictable discourse structure in which the conversational roles shape how

different types of speech relate to perceiver inferences over time. In the dyadic interaction paradigm, conversational roles are more symmetric and fluid. Both speakers are likely to ask questions and share information, and the conversation may be less likely to produce the kind of temporally structured patterns visible in the standard stimulus paradigm. The paradigm-moderated trajectory plots make this distinction vivid: for *inform* dialogue acts, target speech maintains a stable elevation while interactant speech rises to converge with it across the standard stimulus chapter, whereas both remain flat in the dyadic interaction paradigm. For *question* dialogue acts, both speakers show a shared increase in similarity across the standard stimulus chapter that appears to be absent in dyadic interaction studies.

These standard stimulus trajectory shapes relate to the general temporal increase observed in Study 1 in different ways depending on the dialogue act. Study 1 found that perceiver inferences were most semantically similar to speech occurring later in the video chapter, aggregated across all dialogue act types. The *question* dialogue act pattern is consistent with this: *question* utterances later in the chapter showed higher similarity with perceiver inferences, paralleling the overall temporal increase at the level of a specific dialogue act. The *inform* dialogue act pattern is more complex: target *inform* dialogue acts maintained relatively stable alignment throughout the chapter while interactant *inform* dialogue acts increase in similarity closer to the perceiver inference, suggesting that the overall temporal increase may be partly associated with increasing alignment with interactant speech rather than a uniform temporal pattern across all speakers and dialogue acts.

A more speculative interpretation is that the interview structure of the standard stimulus paradigm may drive this pattern because the conversational roles clearly signal who is disclosing and who is prompting. In peer-to-peer “get to know you” conversations, this signal may be

weaker, and perceivers may default to drawing on whatever speech happened most recently (as in the Study 1 temporal increase) rather than selectively aligning with specific dialogue acts. This interpretation is consistent with the finding that the temporal increase from Study 1 was present across both paradigms, while the dialogue-act-specific effects in Study 2 were confined to standard stimulus contexts, though the wide credible intervals in both paradigms counsel against drawing sharp conclusions about the presence or absence of these effects.

The findings from Studies 1 and 2 together paint a picture of the right side of the lens model: what structural and discursive features of speech are most semantically similar to perceiver inferences. What remains unaddressed is whether the cues that are most similar to perceiver inferences are actually valid. Do the utterances that are most aligned with perceiver inferences also correspond to what targets report thinking and feeling? Study 3 turns to this question by applying the same analytic framework to the left side of the lens model, replacing perceiver inferences with target self-reported thoughts as the criterion.

# IV. STUDY 3: PREDICTORS OF TARGET THOUGHT CONTENT

The previous two studies investigated what structural and discursive cues perceivers may use as sources of information when inferring targets' thoughts and feelings. In the context of Brunswik's (1956) lens model (Figure 2), Studies 1 and 2 examined the right-hand side of the lens: the degree to which structural and discursive features of verbal speech predict how perceivers weight that speech when making inferences. However, Studies 1 and 2 did not examine whether these cues actually correspond to what targets were thinking and feeling. That is, they did not yet examine the left-hand side of the lens model.

The final study (Study 3) applies the same analytic techniques used in Studies 1 and 2 to identify not what portions of a social interaction were most similar to perceivers' inferences but rather what portions were most similar to targets' reported thoughts and feelings, and by extension which features of the social interaction may be diagnostic cues that support inferential accuracy. The dependent variable has therefore shifted from cosine similarity between perceiver inferences and speech (Studies 1 and 2) to similarity between targets' reports of their thoughts and feelings and both targets' and interactants' speech (Study 3), mirroring the structure of the lens model directly: Studies 1 and 2 characterized the right path (perceiver cue utilization), and Study 3 characterizes the left path (cue validity).

As Study 3 is explicitly exploratory, results are organized around four research questions that parallel the hypotheses tested in Studies 1 and 2. The two structural research questions (RQ1 and RQ2) parallel the temporal and speaker comparisons explored in Study 1's Hypotheses 1 and 2, and the two discursive research questions (RQ3 and RQ4) parallel Hypotheses 5 and 6 from Study 2.

**Research Question 1**: Does semantic similarity increase as speech occurs closer to the moment of inference?

**Research Question 2**: Do target thoughts/feelings show greater semantic similarity with target speech than with interactant speech?

**Research Question 3:** Do specific dialogue acts (e.g., *inform*, *suggestion*, *question*) uniquely predict greater semantic similarity with target thought content?

**Research Question 4:** Do *question* dialogue acts by interactants precede dialogue acts that are uniquely high in semantic similarity with target thoughts/feelings?

## Method

### *Included Studies*

Target-reported thought content was available for five of the six studies used in Studies 1 and 2, as the target-reported thoughts and feelings for the New Mothers study had not been previously digitized. However, the New Mothers data may not have been included anyway because the data from the remaining standard stimulus paradigm study, the Middle Eastern Men study, failed to meet model convergence criteria (the Bayesian sampler could not settle on stable estimates, yielding unreliable posterior distributions). In retrospect, this is unsurprising: the study was designed to collect many perceiver inferences about the thoughts of relatively few (9) targets. Thus, when I attempted to reverse the nesting structure of these data, there ended up being insufficient data for reliable estimation of target-level GP hyperparameters and random effect variances. With so few reported thoughts and feelings at the target level, the sampler could not adequately characterize the posterior, yielding Rhat values as high as 1.77 and effective sample sizes below 13. Thus, this study was also excluded from the target analyses, and it is

quite likely that the New Mothers study, with only 14 targets, may also have been excluded for the same reason, even if the targets' thoughts and feelings had been previously digitized. The four remaining studies (all dyadic interaction paradigm studies) were retained for target analyses, as their matched perceiver-target designs provide adequate target-level sample sizes for GP estimation.

### *Data Preparation*

The same transcription, utterance segmentation, and temporal indexing pipeline described in Study 1 was used for Study 3. The key difference was the outcome variable. In Studies 1 and 2, the question was how similar each utterance was to what the perceiver wrote about what the target was thinking. In Study 3, this question shifted to how similar each utterance was to what the target reported they were *actually* thinking. Operationally, this meant computing cosine similarity between each utterance's embedding vector and the corresponding target's self-reported thought content embedding (rather than the perceiver's inference embedding). Target-reported thoughts and feelings, like perceiver inferences, were embedded using the all-MiniLM-L6-v2 model within Sentence-Transformers (Reimers & Gurevych, 2019). The cosine between that vector and the vector of the embedding for every utterance in a stimulus video was the operationalization of semantic similarity.

Table 24 provides descriptive statistics for the cosine similarity between target thoughts/feelings and transcribed speech across the four studies included in this analysis. As in Studies 1 and 2, cosine similarity values were generally positive, modest in magnitude, positively skewed, and spanned a range that included near-zero values reflecting individual utterances with little semantic overlap with the target's reported thought content, all the way to highly similar utterances that were almost identical to target-reported thought content.

**Table 24:**

*Study 3: Descriptive Statistics for Cosine Similarity Between Target Thoughts/Feelings and Speech by Study and Speaker Type*

| Study | Target M (SD) | Target Range | Interactant M (SD) | Interactant Range |
|---|---|---|---|---|
| Negotiation | 0.165 (0.148) | [−0.102, 0.836] | 0.163 (0.147) | [−0.089, 0.907] |
| Round Robin (In Person) | 0.119 (0.085) | [−0.164, 0.756] | 0.114 (0.085) | [−0.208, 0.717] |
| Round Robin (Video Mediated) | 0.137 (0.085) | [−0.150, 0.800] | 0.135 (0.085) | [−0.178, 0.793] |
| STEM | 0.182 (0.148) | [−0.171, 0.857] | 0.167 (0.137) | [−0.222, 0.819] |

*Note. M = mean cosine similarity; SD = standard deviation. Target = semantic similarity between target's reported thoughts and target speech utterances. Interactant = semantic similarity between target's reported thoughts and interactant speech utterances.*

### *Bayesian Model Specification*

The Bayesian multilevel model specification for Study 3 mirrors exactly the model used in Study 1, with one change: the dependent variable is the cosine similarity between target-reported thoughts and feelings and transcribed speech, rather than between perceiver inferences and transcribed speech. The model equation, priors, and analytical stages (individual models, univariate meta-analysis, multivariate trajectory synthesis) were identical in form to those described in Study 1 and are not repeated here. As in Studies 1 and 2, DAMSL dialogue act categories were used for the discursive predictors analyses. Utterances were categorized using the same seven-category dialogue act coding scheme as Study 2 (see Table 16).

## Results

Study 3's results are organized into two sections: structural predictors (RQ1 and RQ2, paralleling Study 1) and discursive predictors (RQ3 and RQ4, paralleling Study 2). In all analyses below, "semantic similarity" refers to the cosine similarity between speech utterance

embeddings and target-reported thought embeddings (not perceiver inference embeddings, as in Studies 1 and 2). Higher values indicate that what was said more closely matched what the target reported they were actually thinking and feeling. The structural predictors analyses use the same single-model-per-study approach as Study 1. The discursive analyses use the same approach as Study 2 in which separate models were fit per dialogue act per study and subsequently individual meta-analyses were fit for each individual dialogue act. Because all four studies were dyadic interaction paradigm studies, moderation by paradigm was not possible in either set of analyses.

### *Structural Predictors*

**Stage 1: Individual Bayesian Multilevel Models.** All four models were fit using the same Bayesian multilevel Gaussian process model specification from Study 1. Model convergence was uniformly excellent across all four included studies: maximum Rhat values were 1.00 in all cases and minimum effective sample sizes (ESS) ranged from 1,901.95 (Round Robin, Video Mediated) to 7,080.07 (Negotiation), well above the threshold for reliable posterior inference. Table 25 summarizes convergence statistics for all four models.

**Table 25:**

*Study 3: Bayesian Multilevel Model Convergence Statistics*

| Study | N Utterances | Max R-hat | Min Bulk ESS | Min Tail ESS |
|---|---|---|---|---|
| Negotiation | 3,294 | 1.00 | 4,932.34 | 7,080.07 |
| Round Robin (In Person) | 44,630 | 1.00 | 2,940.15 | 2,837.29 |
| Round Robin (Video Mediated) | 38,367 | 1.00 | 2,738.02 | 1,901.95 |
| STEM | 23,875 | 1.00 | 4,022.37 | 3,754.56 |

Individual study fixed effects (mean cosine similarity between target thoughts and speech) were consistently positive across studies and speaker types, ranging from approximately 0.11 to 0.18 (Table 26). Speaker contrasts at the individual study level showed minimal and

inconsistent differences between target and interactant speech. The mean Target–Interactant difference ranged from 0.00 (Negotiation) to 0.01 (STEM; Round Robin, In Person), with posterior probabilities of Target > Interactant ranging from 0.53 to 0.83. No individual study provided strong evidence that targets' speech was more semantically similar than interactants' speech to the content of targets' reported thoughts.

**Table 26:**

*Study 3: Individual Model Fixed Effect Estimates and Speaker Contrasts with 95% CIs*

| Study | Speaker | *β* [95% CI] | SE | Target-Interactant Δ [95% CI] | P(T > P) |
|---|---|---|---|---|---|
| Negotiation | Target | 0.18 [0.09, 0.27] | 0.04 | 0.00 [−0.12, 0.13] | 0.53 |
| | Interactant | 0.17 [0.09, 0.27] | 0.04 | | |
| Round Robin (In Person) | Target | 0.12 [0.12, 0.12] | 0.00 | 0.01 [−0.03, 0.07] | 0.81 |
| | Interactant | 0.11 [0.05, 0.15] | 0.02 | | |
| Round Robin (Video Mediated) | Target | 0.14 [0.13, 0.14] | 0.00 | 0.00 [−0.01, 0.03] | 0.81 |
| | Interactant | 0.13 [0.12, 0.15] | 0.01 | | |
| STEM | Target | 0.18 [0.14, 0.21] | 0.02 | 0.01 [−0.03, 0.05] | 0.83 |
| | Interactant | 0.17 [0.15, 0.18] | 0.01 | | |

*Note. β = mean cosine similarity. SE = standard error. Δ = Target − Interactant. Pr(T > I) = posterior probability that target speech exceeds interactant speech in semantic similarity to target thoughts. Rhat = 1.00 for all estimates; Bulk ESS ≥ 2,738 for all estimates.*

Gaussian process hyperparameter estimates (Table 27) reflected modest temporal variation in all four studies. Gaussian process amplitude values (how much thought-speech similarity fluctuates over the chapter) were generally very small, ranging from approximately 0.00 (Round Robin, In Person, target) to 0.06 (Negotiation, target), and were on average smaller

than those observed in the perceiver-based models from Study 1. Gaussian process length-scale values (how smoothly those fluctuations unfold) were also small and varied across studies, ranging from 0.11 (Round Robin, Video Mediated, target) to 0.73 (Negotiation, interactant). Posterior trajectories are shown in Figure 25.

***Table 27:***

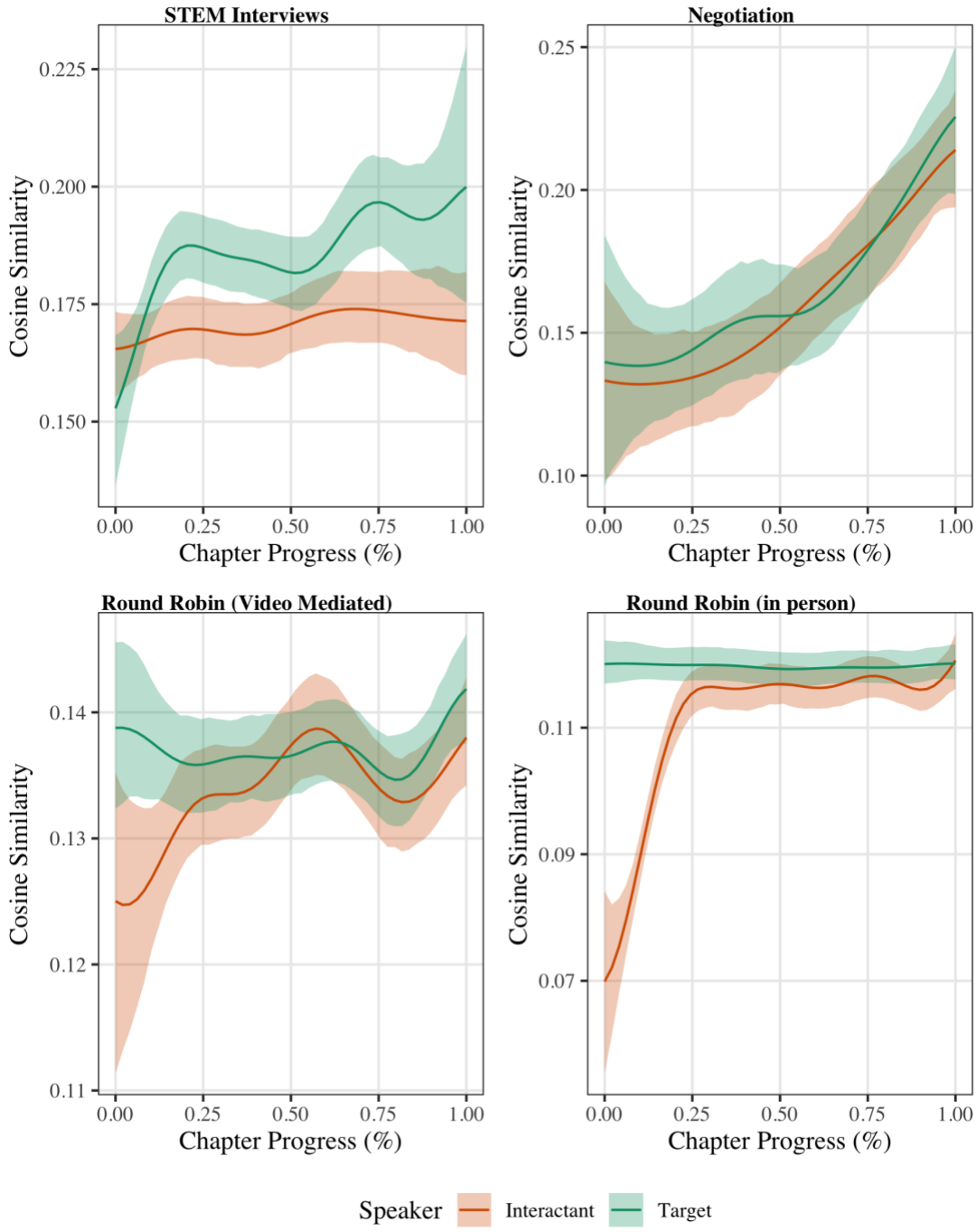
*Study 3: Individual Model Gaussian Process Hyperparameter Estimates with 95% CIs*

| Study | Speaker | Amplitude M (Median, SD) | Amplitude 95% CI | Length-Scale M (Median, SD) | Length-Scale 95% CI (lscale) |
|---|---|---|---|---|---|
| Negotiation | Target | 0.06 (0.05, 0.03) | [0.02, 0.14] | 0.53 (0.40, 0.48) | [0.02, 1.78] |
| | Interactant | 0.05 (0.05, 0.03) | [0.02, 0.14] | 0.73 (0.60, 0.57) | [0.04, 2.20] |
| Round Robin (In Person) | Target | 0.00 (0.00, 0.01) | [0.00, 0.01] | 0.42 (0.04, 3.83) | [0.01, 2.62] |
| | Interactant | 0.03 (0.03, 0.02) | [0.01, 0.09] | 0.26 (0.17, 0.28) | [0.02, 1.02] |
| Round Robin (Video Mediated) | Target | 0.01 (0.00, 0.01) | [0.00, 0.02] | 0.11 (0.04, 0.35) | [0.01, 0.69] |
| | Interactant | 0.01 (0.01, 0.01) | [0.00, 0.04] | 0.21 (0.08, 0.38) | [0.01, 1.18] |
| STEM | Target | 0.03 (0.03, 0.02) | [0.01, 0.08] | 0.15 (0.07, 0.20) | [0.01, 0.73] |
| | Interactant | 0.01 (0.01, 0.01) | [0.00, 0.03] | 0.29 (0.06, 0.79) | [0.01, 2.24] |

*Note. sdgp = GP amplitude (magnitude of temporal variation in cosine similarity). lscale = GP length-scale (smoothness of temporal trajectory). All estimates have Rhat = 1.00.*

**Figure 25:**

*Study 3: Individual Studies' Estimated Semantic Similarity Between Target Thoughts and Speech by Speaker Over Time*



**Stage 2A: Univariate Meta-Analysis of Individual Parameters.** Following the same meta-analytic procedure used in Study 1, I pooled estimates across the four included studies using Bayesian random-effects meta-analyses.

The pooled fixed effect estimates (Table 28) show that target and interactant speech were similarly aligned with target-reported thoughts: $\beta$(Target) = 0.146 [0.10, 0.21] and $\beta$(Interactant) = 0.145 [0.09, 0.20]. In practical terms, what a target said was no closer to what they were thinking than what their conversation partner said, a finding that parallels the null speaker effect from Study 1, which found that perceivers appeared to pay equal attention to target and interactant speech in making their inferences. Pooled Gaussian process amplitude estimates were very small across both speakers: $\alpha$(Target) = 0.021 [0.012, 0.037] and $\alpha$(Interactant) = 0.024 [0.014, 0.042], indicating that temporal variation within chapters was modest after accounting for mean-level similarity. Pooled Gaussian process length-scale estimates were also small, suggesting that whatever temporal variation exists tended to be short-range rather than smooth global trends: $\ell$(Target) = 0.138 [0.088, 0.214] and $\ell$(Interactant) = 0.154 [0.098, 0.242]. Forest plots for all parameters are shown in Figures 26 and 27.

**Table 28:**

*Study 3: Stage 2A Univariate Meta-Analytic Parameter Estimates with 95% CIs*

| Category | Parameter | Pooled [95% CI] | τ [95% CI] |
|---|---|---|---|
| Fixed Effects | β (Target) | 0.146 [0.10, 0.21] | 0.036 [0.010, 0.117] |
| Fixed Effects | β (Interactant) | 0.145 [0.09, 0.20] | 0.033 [0.009, 0.119] |
| GP Hyperparameters | α (Target) | 0.021 [0.012, 0.037] | 0.075 [0.004, 0.246] |
| GP Hyperparameters | α (Interactant) | 0.024 [0.014, 0.042] | 0.070 [0.003, 0.229] |
| GP Hyperparameters | ℓ (Target) | 0.138 [0.088, 0.214] | 0.068 [0.003, 0.224] |
| GP Hyperparameters | ℓ (Interactant) | 0.154 [0.098, 0.242] | 0.067 [0.003, 0.226] |
| Random Effects SD | τ_perceiver (Target) | 0.011 [0.004, 0.028] | 0.285 [0.012, 0.963] |
| Random Effects SD | τ_target (Target) | 0.011 [0.007, 0.023] | 0.236 [0.011, 0.883] |
| Random Effects SD | τ_chapter (Target) | 0.042 [0.029, 0.063] | 0.306 [0.147, 0.780] |
| Residual SD | σ (Residual) | 0.105 [0.074, 0.160] | 0.364 [0.191, 0.816] |

*Note. β = mean cosine similarity (fixed effects). α = GP amplitude (back-transformed from log scale). ℓ = GP length-scale (back-transformed from log scale). τ = between-study heterogeneity SD. Fixed effects on original scale; variance components back-transformed from log scale.*

**Figure 26:**

*Study 3: Stage 2A Forest Plot of Pooled Fixed Effects*

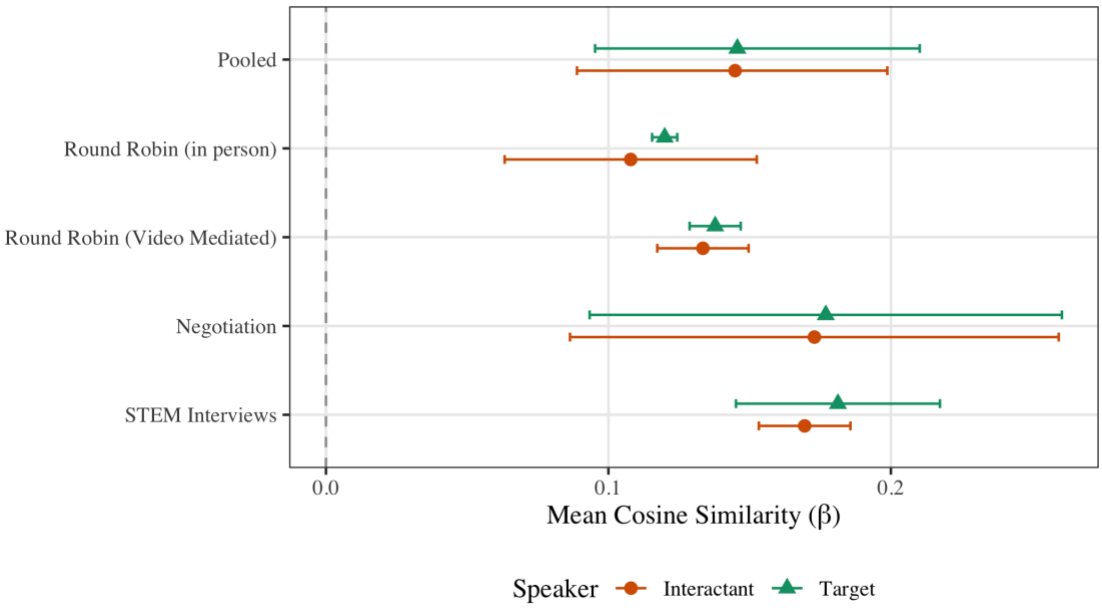
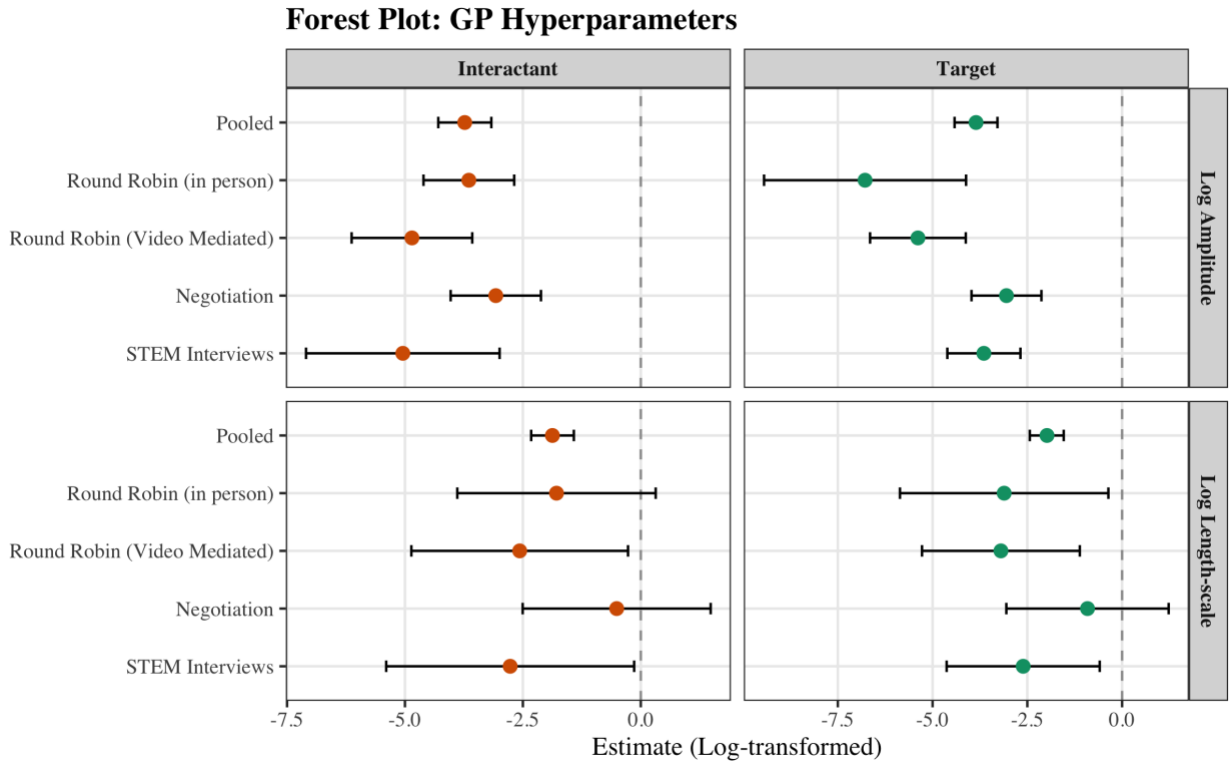


**Figure 27:**

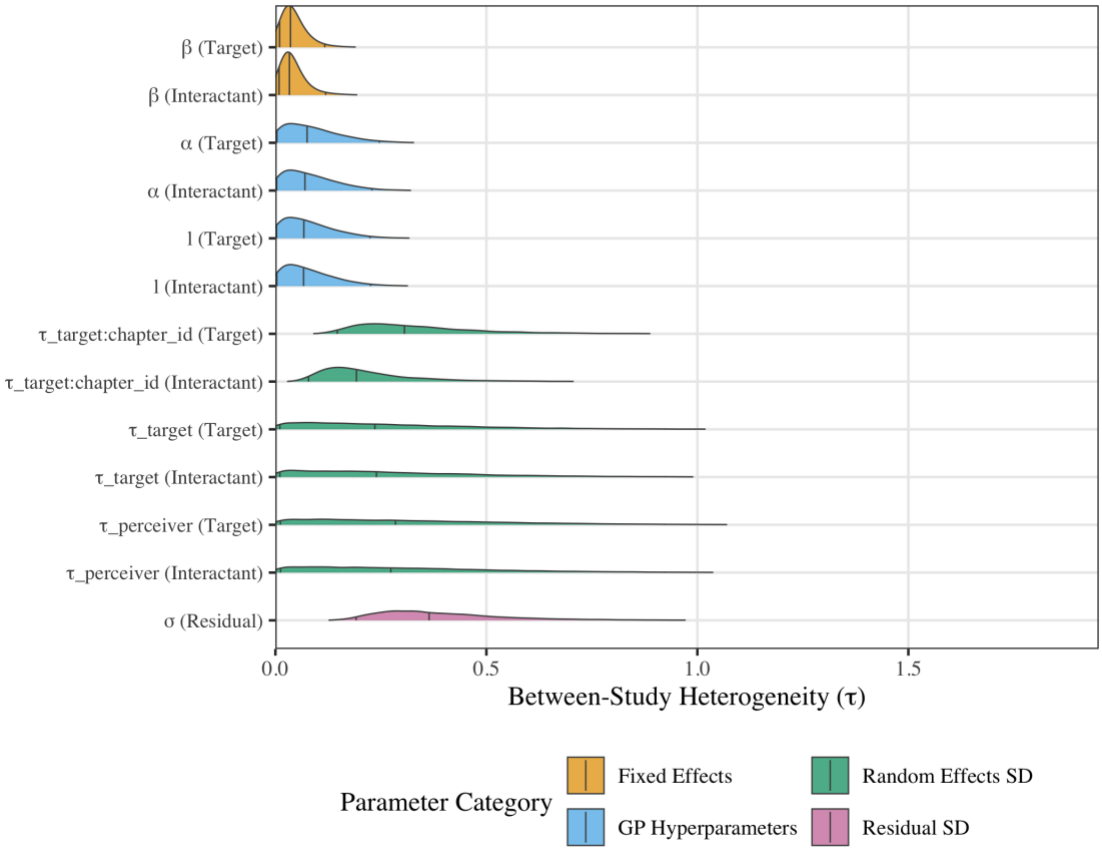
*Study 3: Stage 2A Forest Plot of GP Hyperparameters*




**Between-Study Heterogeneity.** Figure 28 displays the posterior distributions of τ for all meta-analytic parameters in the structural target models. The overall pattern was qualitatively similar to Study 1, but with several differences that are informative about the target-side perspective and the composition of contributing studies.

**Figure 28:**

*Study 3: Between-Study Heterogeneity (τ) Posterior Distributions Across Meta-Analytic Parameters*




Fixed effect heterogeneity was modest: $\tau = 0.036$ [95% CI: 0.010, 0.117] for target speech and $\tau = 0.033$ [95% CI: 0.009, 0.119] for interactant speech. The target-speech heterogeneity is roughly twice the corresponding Study 1 value ($\tau = 0.017$), while the interactant-speech heterogeneity is comparable to Study 1 ($\tau = 0.034$). The doubling of target-speech heterogeneity suggests somewhat greater variability in mean-level thought-speech alignment across the four contributing studies.

The Gaussian process hyperparameter heterogeneity was moderate but (unlike Study 1) nearly symmetric across speakers: amplitude τ values were 0.075 [0.004, 0.246] for target and 0.070 [0.003, 0.229] for interactant, and length-scale τ values were 0.068 [0.003, 0.224] and 0.067 [0.003, 0.226], respectively. The contrast with Study 1 is instructive. In Study 1, interactant amplitude heterogeneity ($\tau = 0.397$) far exceeded target amplitude heterogeneity ($\tau = 0.077$), producing the most prominent asymmetry in the τ ridge plot. In Study 3, this asymmetry is absent, and the likely explanation is straightforward: all four studies in the target analysis used the dyadic interaction paradigm, so the standard stimulus/dyadic interaction contrast that drove the speaker-type heterogeneity in Study 1 is simply not present. This provides indirect support for the interpretation offered in the Study 1 paradigm moderation analyses, namely, that the elevated interactant heterogeneity in Study 1 was attributable to paradigm-related differences in the interactant role rather than to some general property of interactant speech.

Random effect variance components showed the largest between-study variability, as expected. The perceiver and target-level random effect τ values ranged from 0.236 to 0.306, with the chapter-level random effect for target speech showing the largest value ($\tau = 0.306$ [0.147, 0.780]). The residual standard deviation showed $\tau = 0.364$ [95% CI: 0.191, 0.816], indicating that the amount of observation-level variability remaining after accounting for the fixed and random structure differed meaningfully across the four studies. Given the differences in sample size, number of targets, and conversation length across these studies, this level of variance component heterogeneity is not unexpected.

**Stage 2B: Multivariate Synthesis of Pointwise Posterior Trajectories.** Following the same spline-based multivariate synthesis approach as Studies 1 and 2, pointwise posterior trajectories were pooled across the four studies to address RQ1 (temporal increase) and RQ2

(target vs. interactant speech). The same knot configuration (knots at 0.1, 0.5, 0.9) and basis coefficient structure as Studies 1 and 2 were used.

***Research Question 1: Does semantic similarity increase as speech occurs closer to the moment of inference?*** There was strong evidence for a positive temporal trend in target thought–speech cosine similarity over the course of video chapters. The overall endpoint change (Δ) from $t = 0$ *t*o t = 1 was 0.080 [0.024, 0.139], P(+) = 0.996, and corresponded to an evidence ratio of 234.29 (Table 29). This finding directly parallels Hypothesis 1 from Study 1, where temporal increase was also observed for perceiver-based semantic similarity: speech occurring closer to the moment of inference was both more semantically aligned with what perceivers inferred and what targets were actually thinking.

Examining the temporal trend by speaker, interactant speech showed negligibly stronger evidence for temporal increase (endpoint Δ = 0.100 [0.007, 0.200], P(+) = 0.981, ER = 50.61) than target speech (endpoint Δ = 0.060 [−0.000, 0.127], P(+) = 0.974, ER = 37.46). Analyzing the Gaussian processes estimates in segments suggested that the increase was concentrated at the beginning and end of chapters: the second quintile contrast (20–40% vs. 0–20%) showed moderate evidence (P = 0.928), and the final quintile contrast (80–100% vs. 60–80%) also showed moderate evidence (P = 0.942), with comparably less change in the middle portions of the chapter (Q3 vs. Q2: P = 0.736; Q4 vs. Q3: P = 0.571). This pattern of recency somewhat mirrors the general trajectory pattern observed in Study 1 for perceiver inferences.

***Research Question 2: Do target thoughts show higher semantic similarity with target speech than with interactant speech?*** There was no meaningful evidence that targets' self-reported thoughts were more semantically similar to target speech than to interactant speech. This null finding is perhaps counterintuitive: one might expect that what a person says would

more closely match what they are thinking than what their conversation partner says. However, it aligns with the null speaker effect from Study 1, where perceivers also did not differentially weight target speech. On both sides of the lens model, speaker identity appeared to carry little unique signal. The pooled mean-level speaker difference (Target − Interactant) was −0.004 [−0.087, 0.076], P(Target > Interactant) = 0.454. Endpoint and median derivative contrasts were similarly centered on zero: endpoint Δ = −0.041 [−0.160, 0.073], P = 0.230; median derivative Δ = −0.031 [−0.149, 0.101], P = 0.259, suggesting that there were not differing trajectories in the similarity between target or interactant speech and targets' reported thoughts and feelings. The pooled temporal trajectories are shown in Figure 29.

**Table 29:**

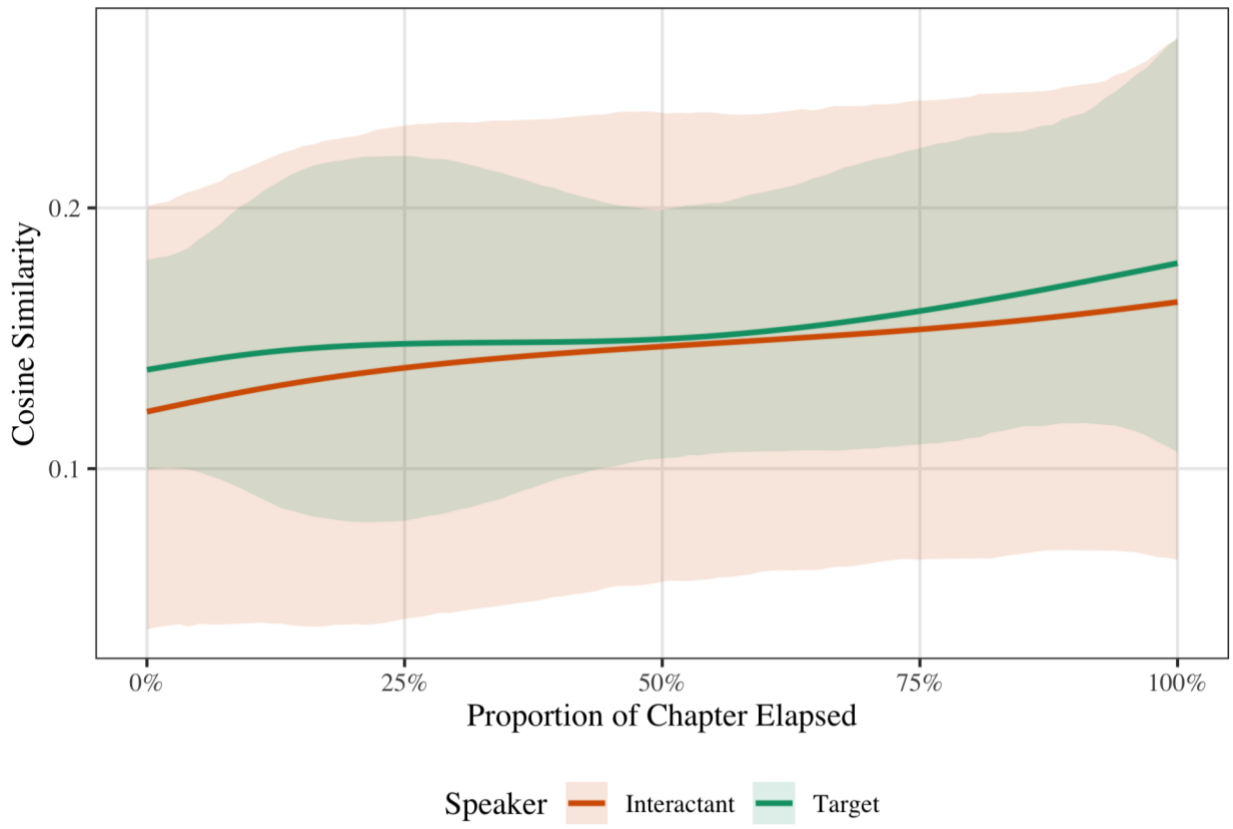
*Study 3: Stage 2B Trajectory Hypothesis Tests with 95% CIs*

| Hypothesis | Method | Speaker | Estimate | 95% CI | P(Dir) |
|---|---|---|---|---|---|
| RQ1 Temporal | Endpoint (Δ t=0 to t=1) | Overall | 0.080 | [0.024, 0.139] | 0.996 |
| RQ1 Temporal | Median Derivative | Overall | 0.061 | [0.005, 0.117] | 0.980 |
| RQ1 Temporal | Endpoint (Δ t=0 to t=1) | Interactant | 0.100 | [0.007, 0.200] | 0.981 |
| RQ1 Temporal | Endpoint (Δ t=0 to t=1) | Target | 0.060 | [−0.000, 0.127] | 0.974 |
| RQ1 Temporal | Seg: Q2 vs Q1 (20–40% vs 0–20%) | Overall | 0.013 | [−0.005, 0.033] | 0.928 |
| RQ1 Temporal | Seg: Q5 vs Q4 (80–100% vs 60–80%) | Overall | 0.025 | [−0.008, 0.059] | 0.942 |
| RQ2 Speaker | Mean Level | T minus P | −0.004 | [−0.087, 0.076] | 0.454 |
| RQ2 Speaker | Endpoint | T minus P | −0.041 | [−0.160, 0.073] | 0.230 |
| RQ2 Speaker | Median Derivative | T minus P | −0.031 | [−0.149, 0.101] | 0.259 |

*Note. RQ = Research Question. Seg = segmented quintile contrast. T minus P = Target minus Interactant. P(Dir) = posterior probability in the hypothesized direction.*

**Figure 29:**

*Study 3: Pooled Temporal Trajectory of Target Thought–Speech Semantic Similarity by Speaker with 95% CIs*




***Discursive Predictors***

The discursive predictors analyses applied the same dialogue-act-stratified modeling approach used in Study 2 to the target thought-content outcome. Rather than examining how perceivers differentially utilized dialogue acts when making inferences (Studies 1 and 2, right-hand side of the lens model), these analyses examined the degree to which each dialogue act category was semantically aligned with what targets were actually thinking (left-hand side of the lens model). Results are organized around Research Questions 3 and 4.

**Stage 1: Individual Models.** I used the same model equation (Equation 3) to estimate the temporal distribution of semantic similarity for each dialogue act across target and interactant speech within each of the four studies. This resulted in 28 individual models (7 dialogue acts × 4 studies). Model convergence was generally adequate, although models of *acknowledgment* dialogue acts for studies with very few observations showed convergence issues (400 < ESS < 700 and 1.02 < Rhat values < 1.25) and were thus removed from meta-analyses. Individual study fixed effects ranged from approximately 0.10 to 0.25. The numbers of each dialogue act within each study are shown in Table 30. Individual model fit statistics for all dialogue act models are shown in Appendix B.

**Table 30:**

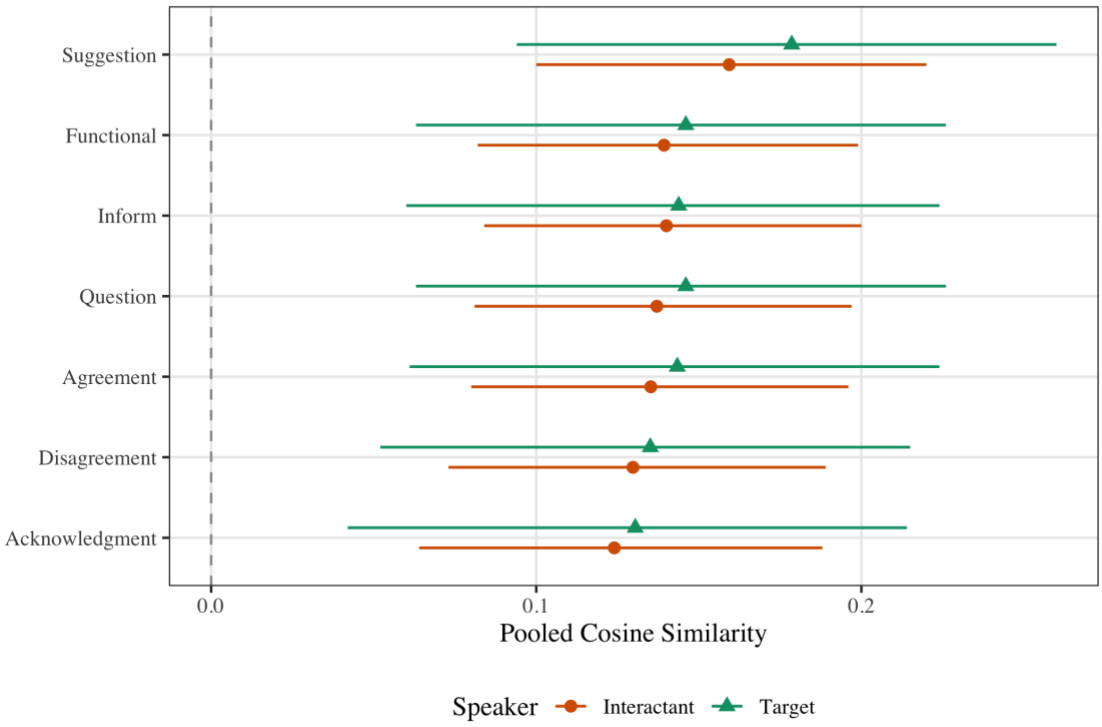
*Study 3: Number of Dialogue Acts by Study*

| Study | Acknowl-edge | Agree | Disagree | Function | Inform | Question | Suggestion |
|---|---|---|---|---|---|---|---|
| Negotiation | 9 | 952 | 258 | 144 | 1,509 | 409 | 13 |
| Round Robin (In Person) | 84 | 9,897 | 5,214 | 1,666 | 20,258 | 7,290 | 221 |
| Round Robin (Video Mediated) | 33 | 9,566 | 4,457 | 1,261 | 17,640 | 5,047 | 302 |
| STEM | 21 | 5,444 | 3,134 | 1,670 | 10,696 | 2,645 | 265 |
| *Total* | *147* | *25,859* | *13,063* | *4,741* | *50,103* | *15,391* | *801* |

**Stage 2A: Univariate Meta-Analyses.** All univariate meta-analysis models achieved adequate convergence (ESS > 1,251, Rhat = 1.00). Pooled fixed effects (Figure 30) showed that mean cosine similarity with target thoughts ranged from 0.112 to 0.181 across dialogue acts and speakers. Focal contrasts comparing *inform*, *suggestion*, and *question* dialogue acts to the baseline acts yielded evidence that, *suggestion* (not *question*) dialogue acts showed the strongest overall semantic similarity with targets' reported thoughts and feelings: target-spoken *suggestions* were 0.038 higher than other dialogue acts (Pr = 0.999), and interactant-spoken *suggestions* were 0.025 higher (Pr = 0.989). Neither *inform* nor *question* dialogue acts showed individual effects. In concrete terms, when a speaker offered advice or a recommendation (e.g., *"You should try applying to graduate school"*), the utterance was more closely aligned with what the target was actually thinking than when a speaker shared information, asked a question, or engaged in any other type of dialogue act. This is the one dialogue act category that showed a clear signal on the cue validity side of the lens model. Speaker contrasts were null across all dialogue acts, mirroring the structural RQ2 finding. The pooled GP hyperparameters did not show evidence for robust speaker effects in either pooled amplitude or length-scale estimates.

**Figure 30:**

*Study 3: Pooled Semantic Similarity and 95% CIs by Dialogue Act*




**Stage 2B: Multivariate Synthesis of Pointwise Posterior Trajectories.** I used the same spline-based multivariate synthesis approach as in the structural predictors to pool pointwise posterior trajectories across the four studies for each dialogue act separately. The same knot configuration (knots at 0.1, 0.5, 0.9)[27] was used for all models. Per-dialogue-act trajectory plots are shown in Figure 31. Temporal trends within individual dialogue acts were generally modest and inconsistent. *Agreement* dialogue acts showed weak evidence for a positive temporal trend (endpoint = 0.009, Pr = 0.863), as did *acknowledgment* dialogue acts (endpoint = 0.023, Pr =

[27] Sensitivity analyses for the random effects structure and knot placement showed negligible sensitivity (maximum $|\Delta| = 0.001$ across all tested configurations), indicating that the trajectory estimates were robust to analytic choices even if the variance ratio suggests some loss of uncertainty information.

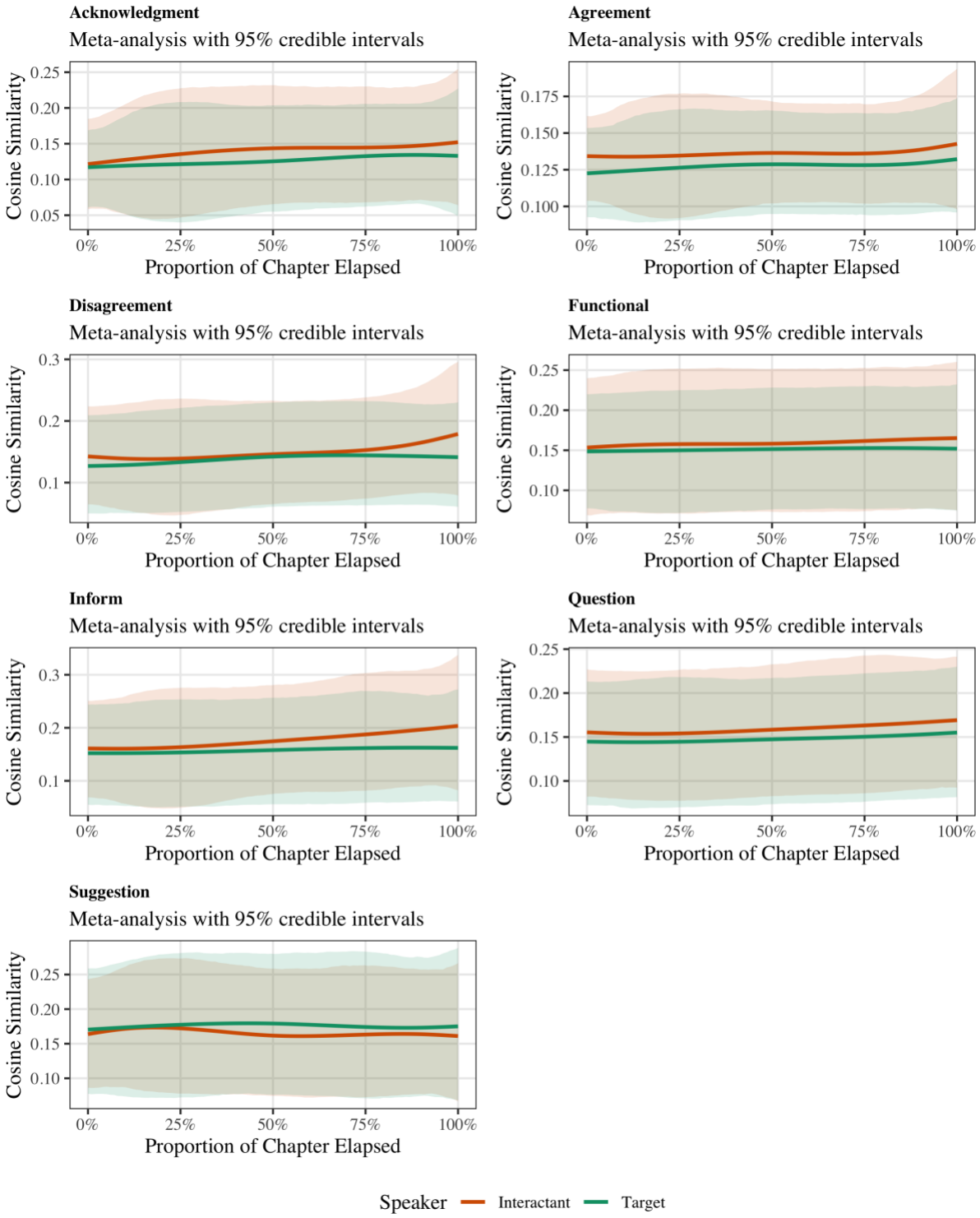
0.837), but no individual dialogue act reached decisive evidence for temporal increase[28]. Speaker contrasts within individual dialogue acts were uniformly null, consistent with the structural RQ2 finding.

The variance ratio diagnostic (which assesses whether the posterior-mean approximation used in the trajectory synthesis adequately captured the full within-study uncertainty) yielded a mean ratio of 2.80 (threshold = 3.0), with only 3 of 10 spline coefficients meeting the adequacy criterion. This indicates that the posterior-mean approximation used in the trajectory synthesis may not fully capture the within-study posterior uncertainty for the discursive models, and combined trajectory estimates should be interpreted with appropriate caution. As no clear (or even suggestive) evidence was identified in the pooled posterior trajectories, this caution is noted for transparency but does not materially affect interpretation of the findings.

[28] Although in Figure 31, *disagreement* may have appeared to increase in semantic similarity towards the end of the chapter, the increased variability in the same time period does not appear robust when examined using the metric of endpoint change.

**Figure 31:**

*Study 3: Pooled Pointwise Posterior Trajectories by Speaker and Dialogue Act*



**Research Question 3 Heterogeneity.** The Approach A fixed effect heterogeneity was generally small (Figure 32), indicating that the mean level of semantic similarity between specific dialogue acts and target-reported thoughts was relatively consistent across the four studies, a pattern that held despite the diversity of conversational contexts, from structured negotiations in the Negotiation study to free-form acquaintanceship conversations in the Round Robin studies. The Approach B per-dialogue-act plots (Figure 33) showed uniformly small heterogeneity across dialogue act categories for the fixed effects, with no individual dialogue act showing markedly elevated τ. Among the Gaussian process hyperparameters, *acknowledgment* and *suggestion* showed somewhat wider τ credible intervals, though this likely reflects the small number of observations for these low-frequency categories (see Table 30) rather than genuinely different temporal dynamics across studies.

**Figure 32:**

*Study 3: RQ3 Between-Study Heterogeneity (τ): Approach A Posterior Distributions*

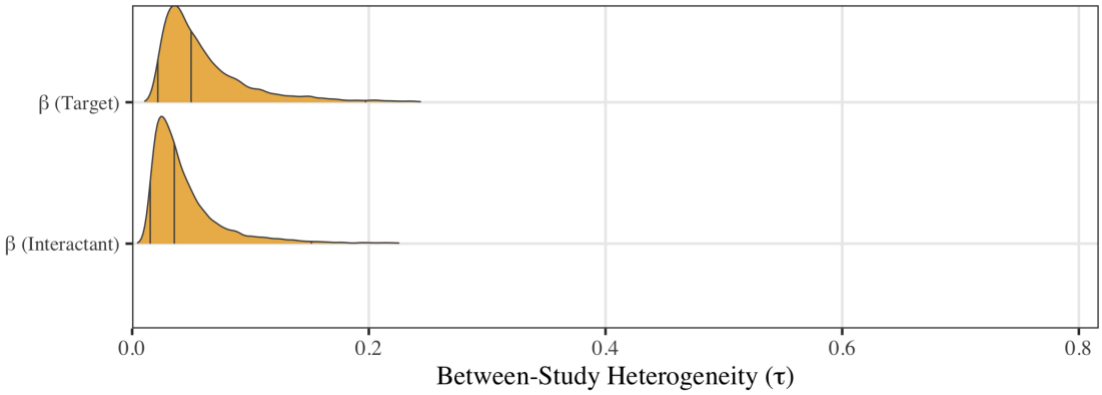
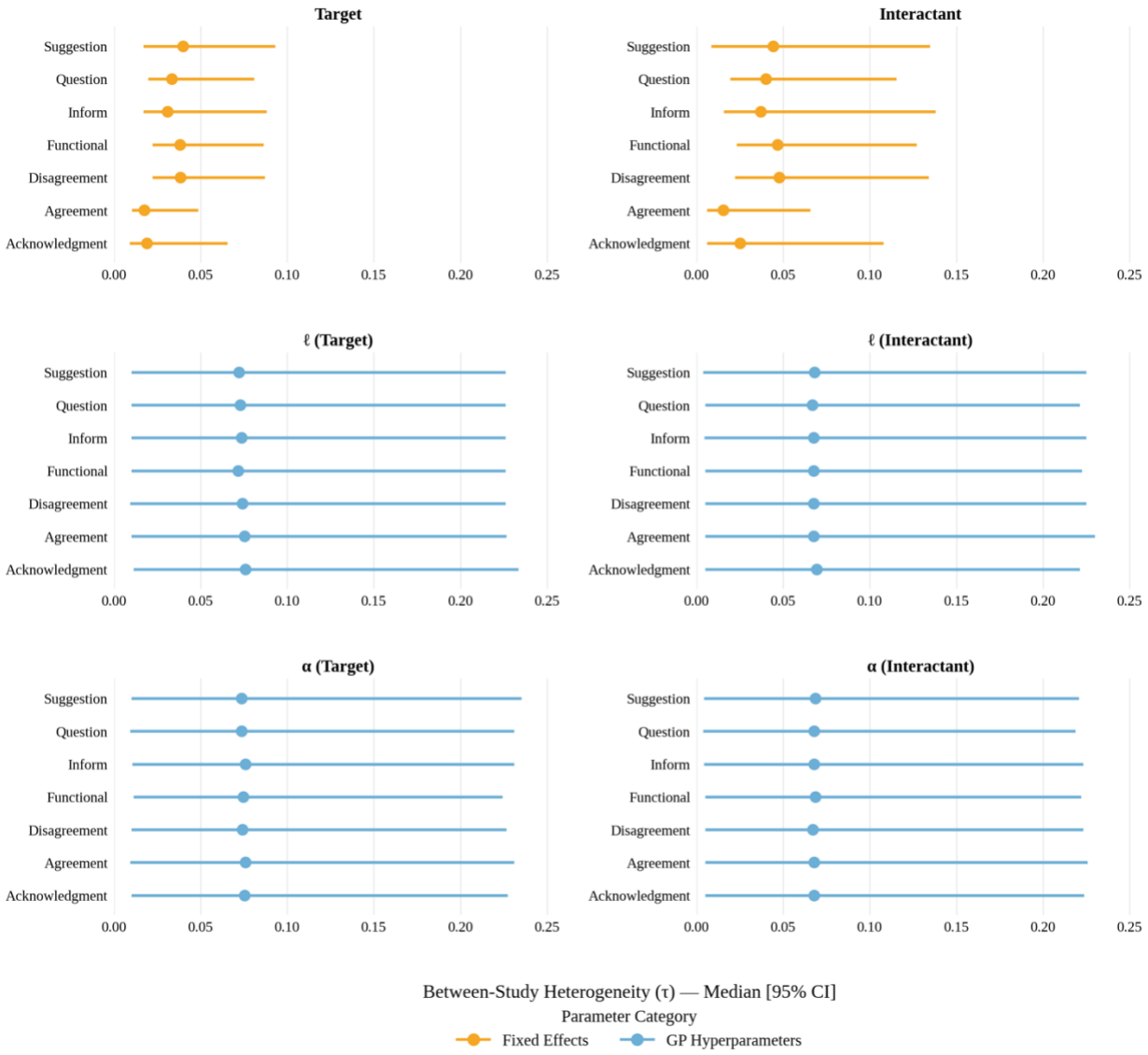


**Figure 33:**

*Study 3: RQ3 Between-Study Heterogeneity (τ) by Dialogue Act: Approach B Point Estimates and 95% CIs*




**Research Question 4: Questions as Cues to Targets' Thoughts and Feelings.** RQ4 examined whether utterances following interactant *question* dialogue acts showed elevated semantic similarity with target-reported thoughts, paralleling H6 from Study 2. This analysis tested the possible importance of *question* dialogue acts as structuring conversational dynamics

by eliciting more thought-reflective responses from conversation partners (e.g., Huang et al., 2017). To examine this, I compared the cosine similarity of utterances immediately following *question* dialogue acts to utterances following all other interactant dialogue acts. *Question* dialogue acts were split into two types: all data labeled "target" refer to target speech following an interactant question; data labeled "interactant" refer to interactant speech following a target question. Two parallel comparisons were thus computed: (1) the focal RQ4 test (target speech following an interactant question vs. target speech following other interactant dialogue acts) and (2) a complementary comparison (interactant speech following a target question vs. interactant speech following other target dialogue acts). The first directly tests whether the post-question target disclosure is more aligned with the target's thoughts; the second tests whether interactant follow-ups to a target's question carry similar signal.

Neither speaker showed meaningful evidence that utterances after *question* dialogue acts were more semantically aligned with target thoughts than utterances following other dialogue acts (Table 31). For target speech, the contrast was essentially zero ($\Delta = -0.001$ [−0.012, 0.010], Pr = 0.571, ER = 1.3), indicating no difference in target thought alignment for target utterances following interactant questions versus those following other interactant acts. For interactant speech, a small positive but inconclusive effect emerged ($\Delta = 0.004$ [−0.007, 0.016], Pr = 0.768, ER = 3.3). The temporal trajectories (Figure 34) similarly showed no systematic divergence between post-*question* and post-other-act utterances over the course of video chapters.

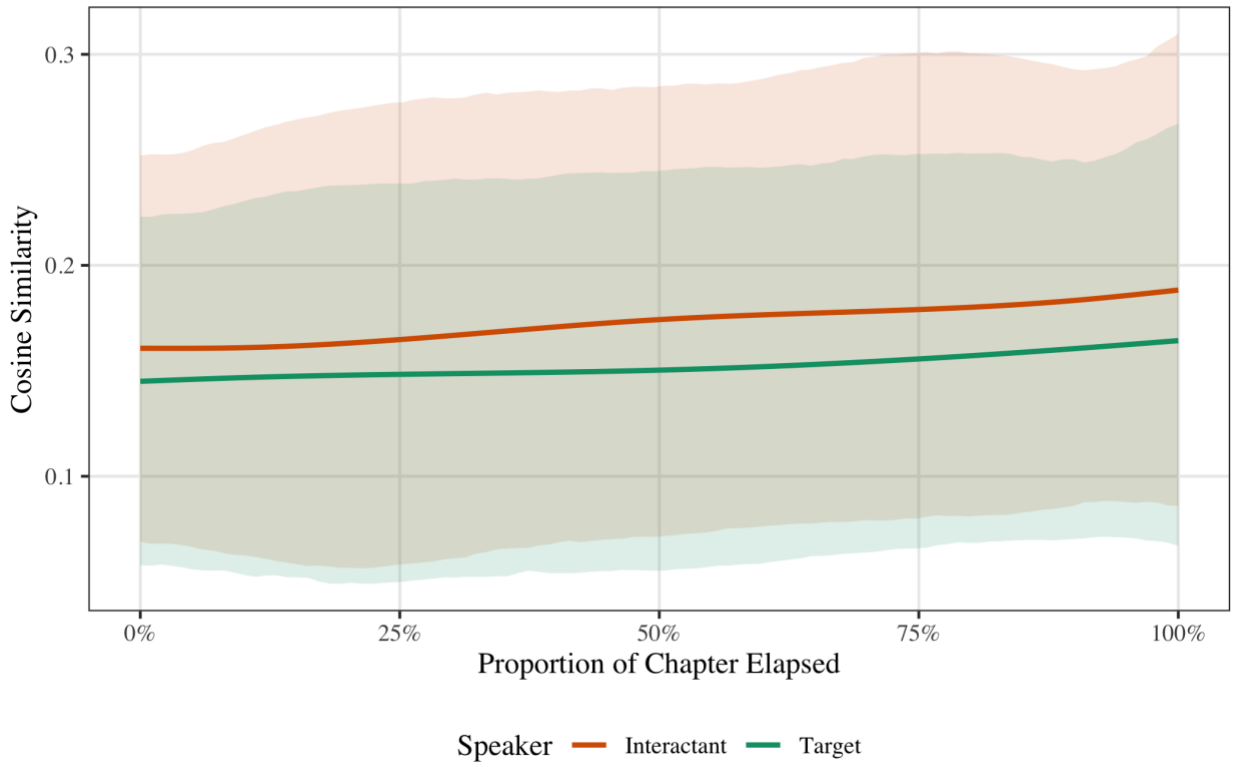
**Table 31:**

*Study 3: Comparison of Utterances Following Questions by Speaker*

| Speaker | Estimate | 95% CI | Pr(Dir.) | Evidence Ratio |
|---|---|---|---|---|
| Target | -0.001 | [−0.012, 0.010] | 0.571 | 1.30 |
| Interactant | 0.004 | [−0.007, 0.016] | 0.768 | 3.30 |

**Figure 34:**

*Study 3: Semantic Similarity of Utterances Following Question Dialogue Acts*

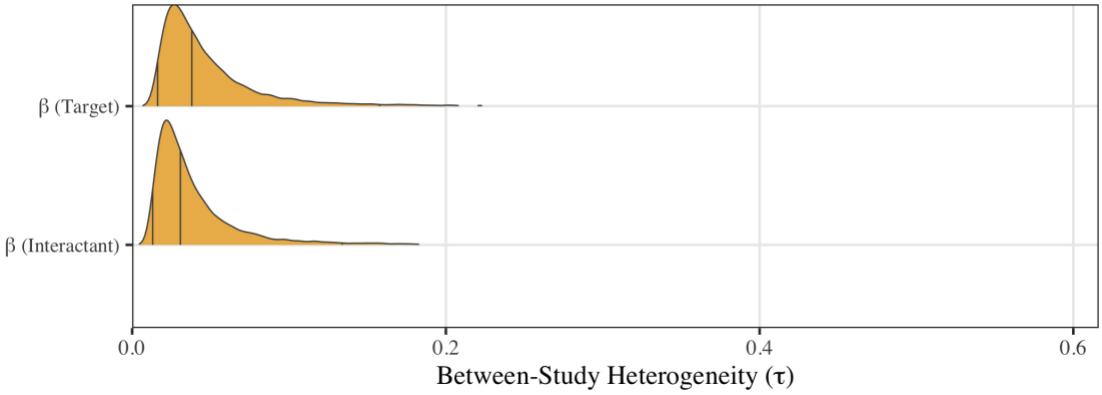



**Research Question 4 Heterogeneity.** The RQ4 heterogeneity patterns largely mirrored those for RQ3. Fixed effect τ values were small and consistent across dialogue acts, and the Approach A GP hyperparameter τ distributions were comparable in magnitude (Figure 35). In

Figure 36, the individual dialogue act plots did not reveal any systematic pattern in which particular dialogue acts showed elevated between-study variability.

**Figure 35:**

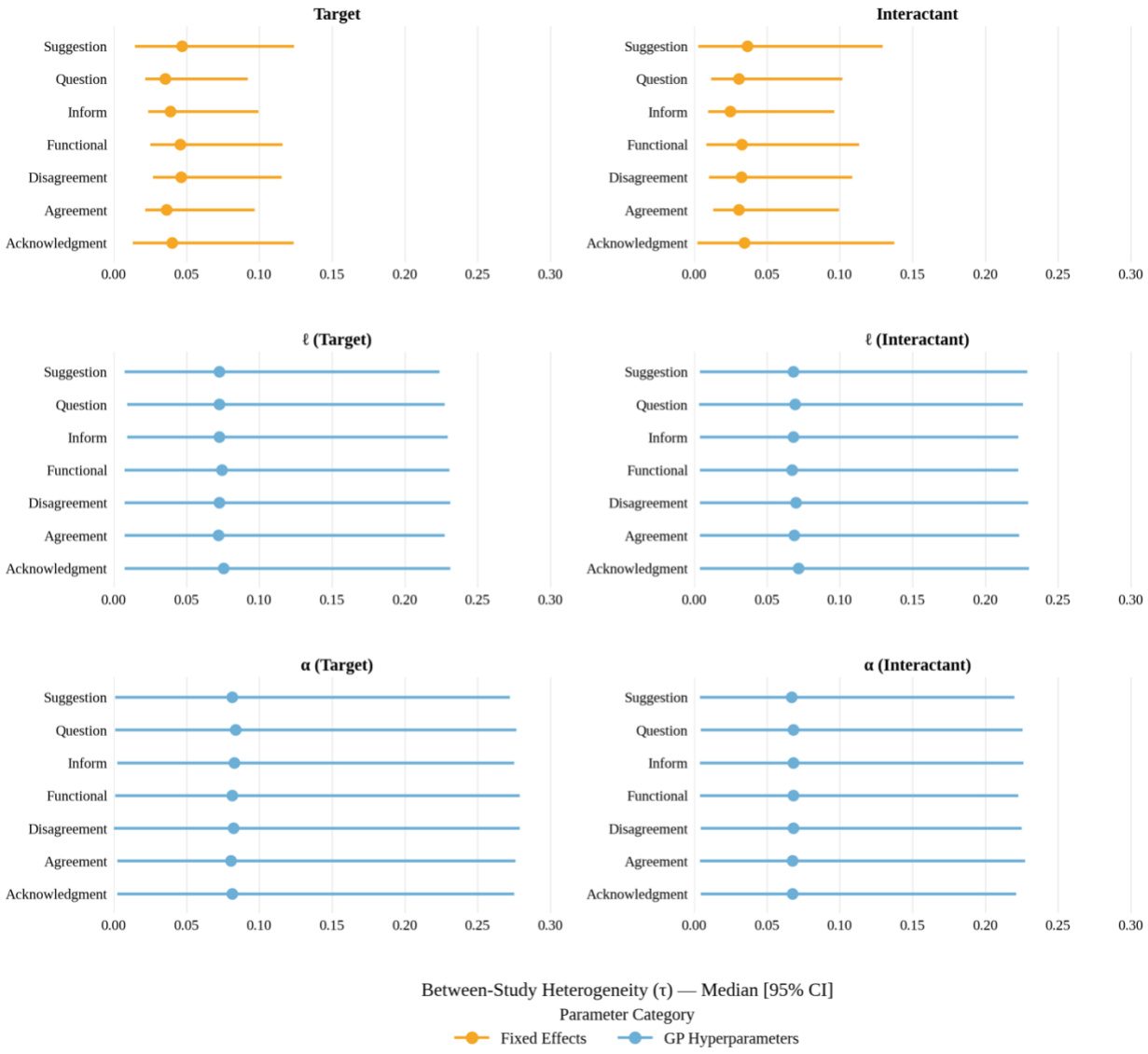
*Study 3: RQ4 Between-Study Heterogeneity (τ): Approach A Posterior Distributions*




One comparison worth drawing is with Study 2, which used the same dialogue-act-stratified approach on the perceiver-inference outcome. The discursive target models showed somewhat smaller between-study heterogeneity overall, particularly for the Gaussian process hyperparameters. Two factors likely contribute: all four contributing studies used the dyadic interaction paradigm, removing the paradigm-related heterogeneity that was present in Study 2; and the target-thought outcome may be inherently less sensitive to study-level contextual features, because the alignment between what someone says and what they are thinking is a more direct relationship than the alignment between what someone says and what an observer infers.

**Figure 36:**

*Study 3: RQ4 Between-Study Heterogeneity (τ) by Dialogue Act: Approach B Point Estimates and 95% CIs*



## Discussion

Study 3 applied the structural and discursive analytic framework of Studies 1 and 2 to the left-hand side of Brunswik's (1956) lens model, replacing perceiver inferences with target-reported thoughts as the comparison criterion. This shift in dependent variable transformed the analysis from one of cue utilization (do perceiver inferences align with what was said?) to one of cue validity (do target thoughts align with what was said?). The juxtaposition of these two analyses is what makes the lens model framework informative: it allows identification of not only what perceivers do but whether what they do is well-calibrated to the actual structure of the cue space.

I summarize the four research questions briefly here and reserve the full cross-study synthesis for the General Discussion (Chapter V), where the convergences and divergences across both sides of the lens are examined together.

**RQ1 (Temporal Increase in Cue Validity)**: Speech occurring closer to the inference point was more semantically similar to what targets reported thinking. This temporal effect was decisive and, if anything, slightly stronger than the analogous finding from Study 1's dyadic-interaction subset on the perceiver side, the most directly comparable Study 1 estimate given that Study 3 includes only dyadic-interaction studies.

**RQ2 (Speaker Differences in Cue Validity)**: Target-reported thoughts were no more semantically similar to their own speech than to the speech of the other interactant. This null finding converged with Hypothesis 2 from Study 1.

**RQ3 (Discursive Cue Validity)**: *Suggestion* dialogue acts showed the strongest alignment with target thought content of any individual dialogue act category, and this held across both target-spoken and interactant-spoken *suggestions*. This finding is notable because

perceivers did not differentially align with *suggestions* in Study 2, creating a specific misalignment between cue validity and cue utilization that will be explored in the General Discussion. Whereas the Study 2 trajectory synthesis provided suggestive evidence that the discursive effects perceivers may be most sensitive to involve *inform* and *question* dialogue acts with temporally distributed patterns in the standard stimulus paradigm, *suggestion* dialogue acts, found to be the most valid cue in Study 3, were not differentiated from other dialogue acts in Study 2.

**RQ4 (Question-Elicitation Effect)**: No evidence emerged that utterances following interactant *question* dialogue acts were more semantically similar to target thoughts. This null finding directly paralleled Hypothesis 6 from Study 2, where interactant *question* dialogue acts showed no unique effect on speech-inference similarity in either paradigm or any temporal region of the chapter.

One important limitation of Study 3 is that the standard stimulus paradigm studies could not be included (there were no digitized target-reported-thought data available for the New Mothers study, and the Middle Eastern Men study had an insufficient number of targets for models to converge). Consequently, Study 3 findings are limited to dyadic interaction paradigm studies, and the paradigm-specific effects identified in Studies 1 and 2 could not be tested on the target side. This constraint is addressed in the "future directions" section of the General Discussion (Chapter V), along with the full interpretation of these and other findings, and their integration with the cue utilization analyses from Studies 1 and 2.

# V. GENERAL DISCUSSION

The three studies in this dissertation represent a novel application of Brunswik's (1956) lens model to inferential accuracy. Using sentence embeddings to compute the cosine similarity between individual utterances and either perceiver inferences (Studies 1 and 2) or target-reported thoughts (Study 3), I have partially mapped both sides of the lens for verbal information as a cue space. Before synthesizing these findings, it is worth being explicit about what the choice of semantic similarity as an outcome variable can and cannot tell us.

## Measurement and Inference

The primary outcome across all three studies is cosine similarity: the degree of semantic overlap between the embedding of a spoken utterance and the embedding of a perceiver's written inference (Studies 1 and 2) or a target's written thought report (Study 3). Higher cosine similarity means that the words spoken and the words written to describe a thought or inference are more semantically alike. This is a measure of similarity, not a direct measure of attention, cognitive processing, or causal influence.

When I describe findings using the lens model terminology of "cue utilization" (Studies 1 and 2) and "cue validity" (Study 3), I am applying a theoretical framework to the observed similarity patterns. The framework is appropriate because it organizes the findings in a way that reveals calibration and miscalibration between what perceivers do and what the environment offers. However, the empirical link between similarity and utilization rests on the assumption that when perceiver inferences are more semantically similar to a particular utterance, perceivers were, in some sense, drawing on that utterance when forming their inference. This assumption is plausible: If a perceiver writes "She was feeling nervous about the interview," and the target said something closely related moments earlier, it is reasonable to think the perceiver was influenced

by that speech. But it is not the only explanation. Perceivers and targets may independently converge on similar language because they share a common conversational context, common cultural scripts, or common expectations about what people think in certain situations.

Throughout this discussion, I describe the data in terms of what was measured (semantic similarity) and will flag when I am making the inferential step from similarity to utilization or validity. Where I use phrases like "perceivers appear to weight recent speech more heavily," the reader should understand this as shorthand for "perceiver inferences are more semantically similar to recent speech, and if we interpret this similarity as reflecting how perceivers constructed their inferences, they appear to more heavily weight recent speech." With this caveat in place, I turn to the specific findings.

**Temporal Increase: The Most Consistent Finding Across the Lens**

The most robust result across all three studies was a temporal increase. Perceiver inferences were more semantically similar to speech occurring closer to the end of a video chapter (Study 1), and target-reported thoughts were also more semantically similar to recent speech (Study 3). This convergence across both sides of the lens model is a central structural finding of this dissertation.

On the perceiver side (Study 1), the synthesized trajectory revealed a gradual, distributed temporal increase in semantic similarity across the chapter. On the target side (Study 3), the same temporal pattern emerged with, if anything, slightly greater strength than the analogous estimate from Study 1's dyadic-interaction subset (the most directly comparable Study 1 estimate, given that Study 3 includes only dyadic-interaction studies). Speech, by either targets *or* interactants, near the end of the chapter was more semantically similar to what targets reported thinking and feeling. This convergence across both sides of the lens has an important

implication for the concern raised in the Study 1 discussion, that the temporal increase might reflect a simple response bias (perceivers parroting recent speech rather than trying to infer target thoughts). If the temporal effect on the perceiver side were solely a response artifact, it would be unlikely to appear on the target side, where the comparison is between speech and what targets themselves reported thinking. The fact that recent speech is also the most semantically similar to target thoughts suggests that the temporal effect may reflect genuine signal in the conversational stream, not merely a perceiver heuristic.

However, an important qualification applies: although the temporal convergence across both sides of the lens is consistent with perceivers being well-calibrated, it does not establish that perceivers are *attending to* recent speech because it is diagnostic. An alternative account is that the convergence is coincidental: perceivers may default to recent speech for reasons unrelated to its diagnostic value (e.g., memory accessibility), and recent speech may happen to be more diagnostic for reasons unrelated to perceiver attention (e.g., conversations naturally build toward the topics that will be on a target's mind at the chapter's end, or targets use recent speech to remind themselves of what they were thinking). The similarity data alone cannot distinguish calibrated attention from spurious coincidence.

Across studies, the strength and shape of the temporal effect differed considerably across conversations. Some produced steep recency gradients while others were essentially flat, and the between-study variability in temporal dynamics was substantial. The pooled trajectory, with its wide credible intervals, is best understood as an average tendency superimposed on a heterogeneous landscape of conversational dynamics.

**Speaker Effects: A Calibrated Non-Difference**

On both sides of the lens model, there was no evidence that target speech occupied a privileged role relative to interactant speech. In Study 1, the contrast between target and interactant similarity was centered near zero across the full trajectory. In Study 3, target-reported thoughts were no more semantically similar to targets' own speech than to their interactants' speech.

This convergent null finding may be more informative than it first appears upon considering possible alternative explanations. If perceiver inferences had been more similar to target speech while target thoughts were equally reflected in both speakers' utterances, this would have indicated miscalibration: perceivers privileging a cue that was not uniquely valid. Conversely, if target thoughts had been more similar to target speech but perceivers failed to detect this asymmetry, a different kind of miscalibration would have been evident. Instead, both sides of the lens show the same pattern: speaker identity does not predict semantic similarity with either inferences or thoughts. If this is interpreted as reflecting perceiver strategy, perceivers are at least trivially calibrated on the speaker dimension: they do not overweight target speech, and target speech does not carry a unique signal that they are missing.

This finding is consistent with a view of dyadic conversations as a site for the *co-construction of understanding* (Clark, 1996; Clark & Brennan, 1991; Schilbach et al., 2013). In this view, both partners arrive at shared meaning together rather than one passively decoding the other. When in conversation, the target is likely to be thinking about what the interactant is saying at least some of the time, and what one person says is itself shaped by what the other has just said (Bavelas et al., 2000). The resulting conversational flow carries information relevant to what a target is thinking regardless of who is speaking, because, in a meaningful sense, both

speakers have a hand in producing those thoughts. That interactant speech is just as semantically similar to target thoughts as target speech is to target thoughts is consistent with this view: target thought content does not appear to be determined by any single speaker's utterances, but is distributed across the interaction as a whole, consistent with dyadic accounts of shared reality in which inner states are held jointly rather than individually (Rossignac-Milon et al., 2021).

An important caveat: the null finding for a speaker effect was limited to studies using the dyadic interaction paradigm, where targets and interactants had symmetric conversational roles. Study 1's exploratory paradigm-moderated analyses (which followed the findings of Hypothesis 4) revealed that interactant speech in the standard stimulus paradigm showed substantially greater temporal variation than target speech in its similarity to perceiver inferences, suggesting that when an interviewer speaks, the content is sometimes highly relevant and sometimes not. This paradigm effect could not be tested on the target side because neither of the standard-stimulus studies could be used in Study 3, a limitation addressed in the Future Directions section below.

**Discursive Cue: Suggestions**

One of the clearest findings in this investigation emerged from the discursive analyses. In Study 3, *suggestion* dialogue acts showed the strongest semantic similarity with target thought content of any individual dialogue act category, with decisive evidence for target-spoken *suggestions* and very strong evidence for interactant-spoken *suggestions*. No other individual dialogue act, including *inform* acts (which share the property of being substantive and information-rich), showed comparable effects.

On the perceiver side (Study 2), however, *suggestion* dialogue acts were not noticeably different from any other dialogue act. No method of assessing differences in speech-inference

similarity across dialogue acts indicated that perceivers differentially aligned with *suggestion* utterances. Additionally, insufficient instances of *suggestion* dialogue acts meant Study 2 was only able to examine the role of *suggestions* in dyadic interaction paradigm studies – given the considerable variability observed in Study 1 between the two paradigms, the role of *suggestions* in standard stimulus studies is a clear next step in this line of research. The contrasts in Study 2 were primarily associated with *inform* and *question* acts (and only in specific paradigm contexts), not by *suggestions*. For inform dialogue acts, target speech maintained a consistently elevated level of similarity across the standard stimulus chapter while interactant speech rose across the chapter steadily to converge with target speech by the end of the chapter. For *question* dialogue acts, both speakers showed a shared temporal pattern in which similarity with perceiver inferences increased across the chapter. However, neither of these temporally distributed patterns extended to *suggestion* dialogue acts at any point in any paradigm, and the wide credible intervals surrounding the *inform* and *question* trajectories mean that even these patterns should be interpreted cautiously.

This creates a specific point of possible misalignment in the lens: the dialogue act category that is most semantically similar to target thought content (*suggestion*) is not the category that is most semantically similar to perceiver inferences. Interpreting this through the lens model framework, if similarity reflects utilization and validity respectively, then the most ecologically valid cue for target thought content is the one perceivers fail to utilize.

One possible explanation for why *suggestions* are particularly reflective of target thought content is that *suggestions* represent moments of active cognitive elaboration. Unlike *agreements* or *acknowledgments*, which are largely reactive, and unlike *inform* dialogue acts, which may convey factual content tangential to the speaker's current cognitive state, *suggestions* require the

speaker to formulate and express a novel thought proposing an action, idea, or interpretation that may directly correspond to what they are thinking (Jourard, 1971). When a speaker makes a *suggestion*, they may be externalizing their current thinking in a way that produces higher semantic overlap with their concurrently reported thoughts. An alternative, though not mutually exclusive, interpretation is methodological: the embedding model may capture more semantic content of *suggestions* because suggestions tend to be longer and more substantive than *acknowledgments* or *agreements*. However, this explanation would also predict elevated effects for *inform* acts, which are similarly substantive yet showed no comparable effect.

So, why might perceivers not align with this cue? Several possibilities merit consideration. First, *suggestions* are relatively rare in the conversations studied here, and their low base rate may reduce their salience to perceivers. Second, perceivers may not experience *suggestions* as more revealing of a speaker's inner state than other dialogue acts, even though the embedding analysis shows higher semantic similarity to target thoughts. Third, perceivers may rely on structural heuristics (recency, temporal position) rather than fine-grained dialogue act discrimination, particularly under the cognitive load of watching and evaluating conversations in real time. The comparatively strong temporal recency effect in perceiver inferences is consistent with such a heuristic: rather than tracking which dialogue acts are most semantically diagnostic, perceivers may default to weighting recent speech regardless of its discursive function.

However, caution is warranted. As this is the first application of the DAMSL taxonomy to inferential accuracy research, it is possible that the dialogue act categories are too broad to capture the distinctions that matter: perhaps a subset of *inform* acts that reflect acute self-disclosure would show high similarity to target thoughts, but this effect was washed out when all *inform* acts were combined in this study. The misalignment between *suggestion* validity and

perceiver utilization could therefore reflect genuine perceiver insensitivity, or it could reflect a mismatch between the coding scheme and the psychological constructs of interest.

**Discursive Cue: Questions**

Surprisingly, the null finding for *question* dialogue acts was consistent across both sides of the lens. In Study 2, Hypothesis 6, speech following *questions* showed no unique effect on speech-inference similarity. In Study 3, Research Question 4, utterances following *questions* also showed no elevated similarity with target thoughts relative to utterances following other dialogue acts. Both contrasts were essentially zero.

This is curious because *questions* are frequently theorized to play a key role in social interaction by eliciting disclosure or structuring conversational dynamics (e.g., Huang et al., 2017). The current findings suggest that, at least as operationalized through the collapsed DAMSL *question* category, *questions* in the dyadic interaction paradigm do not uniquely structure either the similarity between speech and perceiver inferences or the similarity between speech and target thoughts. Interactant *questions* do not appear to elicit responses that are particularly semantically similar to what targets are thinking, nor do perceiver inferences show elevated similarity with post-*question* speech.

One important qualification is that the DAMSL *question* category is broad, collapsing open-ended questions (*"What do you think about that?"*) and closed-ended questions (*"Is that right?"*) into a single category. Open-ended questions may elicit more thought-reflective responses, but this effect could be washed out when all *question* types are combined. Additionally, Study 2 identified paradigm-specific *question* effects that could not be tested in Study 3. In the standard stimulus paradigm, *questions* from both targets and interactants showed increasing similarity with perceiver inferences across the video chapter, with target questions

reaching the highest level by the chapter's end. Both trajectories rose in a roughly parallel fashion, producing a shared temporal pattern that was driven entirely by the two standard stimulus studies, though the credible intervals were notably wide, reflecting considerable between-study heterogeneity. These studies could not be included in Study 3, leaving open the possibility that *questions* play a different role in the standard stimulus paradigm, where interactants act as interviewers and *questions* are structurally more prominent.

A second possibility is that the theoretical claim about *question*-elicited disclosure operates at a different level of analysis than the one examined here. *Questions* may structure the overall trajectory of a conversation without necessarily making the immediately subsequent utterance more thought-reflective. The current analysis compared utterances immediately following a *question* to utterances following other dialogue acts; a longer-range analysis might reveal effects that the utterance-level approach cannot detect (e.g., interactions that as a whole have more *questions* may lead to greater inferential accuracy).

**Pooled Effects Versus Temporal Trajectories**

A pattern that cuts across the findings of all three studies is a distinction between how structural and discursive predictors behave at the level of pooled main effects versus synthesized pointwise posterior trajectories. Structural predictors (e.g., temporal position, speaker identity) were consistent at both levels of analysis across all three studies: the temporal effect was visible in both the meta-analytic fixed effects and the trajectory synthesis, and the null speaker effect (i.e., no difference between target speech and interactant speech) held at both the average level and across the full trajectory. The discursive predictors present a more layered picture.

On the perceiver side (Study 2), *inform* and *question* dialogue acts showed temporally distributed patterns in the standard stimulus paradigm: Target *inform* dialogue acts maintained an

elevated level of similarity while interactant *inform* dialogue acts increased in semantic similarity across a chapter to converge with the target levels, and both speakers' *question* dialogue acts showed increasing semantic similarity with perceiver inferences across the video chapter. However, these effects were paradigm-specific rather than universal, and the wide credible intervals surrounding each trajectory temper the strength of these conclusions. No mean-level differences in speech-inference similarity were observed across dialogue act categories.

On the target side (Study 3), the pattern was nearly the reverse: the most consequential discursive finding, that *suggestion* dialogue acts showed the strongest semantic similarity with target thought content, was a pooled main effect. The temporal patterns within individual dialogue acts were inconsistent and no dialogue act showed decisive evidence for a time-varying relationship with target thoughts.

Thus, when a speaker makes a *suggestion*, the content of that *suggestion* is, on average, more semantically similar to what the target reports thinking, but this elevated similarity does not vary systematically over the course of the conversation in the way that the structural temporal effect does. By contrast, the perceiver-side discursive effects in Study 2, while temporally more variable, were confined to specific dialogue act categories within specific paradigm contexts and were accompanied by substantial uncertainty. Together, these patterns suggest that perceivers' sensitivity to discursive features may be narrower and more context-dependent than their sensitivity to structural features like the temporal increase. If semantic similarity is viewed as reflective of cue utilization and validity, the practical implication is that perceivers may benefit from attending to the discursive content of *suggestions* as a complement to the structural heuristic of recency, rather than relying on structural features alone.

**Between-Study Heterogeneity**

Heterogeneity in average semantic similarity was small in every study, indicating that the fundamental rates of similarity (positive semantic similarity between speaker utterances and either perceiver inferences or target thoughts) replicated across the six datasets regardless of paradigm, topic, or sample. This consistency is encouraging, because it suggests that the average-level alignment is robust enough to survive aggregation across diverse conversational paradigms.

However, heterogeneity in the Gaussian process hyperparameters told a more nuanced story. The largest between-study variability observed anywhere in the dissertation was for interactant amplitude in Study 1, and the comparison with Study 3 (where the same parameter showed far less heterogeneity) helps pinpoint the source: the standard stimulus/dyadic interaction contrast. When the paradigm comparison cannot be made (as in Study 3, where only dyadic interaction studies were included), speaker-type asymmetries in temporal dynamics largely disappear. This is consistent with the argument developed in the Study 1 discussion that the elevated temporal variability for interactant speech in the standard stimulus paradigm reflects the distinctive interviewer role, not a general property of interactant speech.

Variance component heterogeneity was consistently the largest across all studies. This is not unexpected: Pooled averages tend to be more homogeneous than pooled variances in essentially all meta-analytic settings because variance parameters are more sensitive to distributional characteristics that differ across studies. Thus, the prediction intervals implied by these distributions are worth bearing in mind when considering the generalizability of the pooled estimates.

**Broader Implications**

The lens model synthesis revealed an asymmetry between structural and discursive calibration that has both theoretical and practical significance. This asymmetry suggests that there may be room for improvement in perceiver accuracy. If the finding for *suggestion* dialogue acts replicates and if the similarity patterns reflect genuine attentional processes, then the misalignment identifies a specific, actionable target: the dialogue act category most semantically similar to target thoughts is not currently optimally reflected in perceiver inferences. Training perceivers to attend to *suggestions* could, in principle, improve inferential accuracy by redirecting attention toward a valid but underutilized cue. However, this recommendation is conditional on two unresolved questions: whether the *suggestion*-thought similarity reflects something psychologically meaningful (rather than an embedding model artifact) and whether perceivers can be trained to discriminate and attend to different dialogue act types in real time.

More broadly, the findings may speak to the nature of verbal information as a cue in social cognition. Prior work has established verbal information as a robust – perhaps the single strongest – predictor of inferential accuracy (Gesn & Ickes, 1999; Hall & Schmid Mast, 2007; Hodges & Kezer, 2021; Kraus, 2017), and the present investigation adds a unique refinement: there are specific structural and discursive features of verbal information that predict alignment between speech and inferences of thoughts. The temporal position of an utterance, the paradigm context in which it occurs, and (on the target side) whether it occurs as part of a speaker's *suggestion* all correspond to the degree of semantic similarity. This investigation has thus begun to fill in some additional details about the finding that "words matter for inferential accuracy": going from a monolithic claim toward a differentiated map of which words, spoken when, and in what conversational context, carry the most signal.

## Limitations and Future Directions

### *The Standard Stimulus Paradigm Gap*

The inability to include standard stimulus paradigm studies in Study 3 is the most consequential limitation for this investigation. Two of the most interesting findings from Studies 1 and 2 were specific to standard stimulus studies: the paradigm-moderated amplitude and length-scale differences in temporal dynamics, and the distinctive role of interactant-as-interviewer in shaping how speech aligns with perceiver inferences. Yet the left-hand side of the lens could not be mapped for these paradigms, leaving the question of whether the standard stimulus findings on the perceiver side correspond to genuine cue validity entirely unanswered. Further, the most interesting dialogue act finding, that *suggestion* dialogue acts seem to be valid cues that are under-utilized by perceivers, is limited to just the dyadic-interaction paradigms because the standard stimulus paradigm studies lacked sufficient instances of this dialogue act to be modeled.

Addressing this gap requires collecting a substantial dataset using a standard stimulus paradigm with far more targets than is typical of these designs. Standard stimulus paradigm studies have historically been valued precisely because they require a relatively small number of targets: these individuals are filmed in an interview, and those recordings serve as the stimulus set for a larger sample of perceivers. This benefits researchers by minimizing the noise from heterogeneous dyads. It also gives researchers the ability to work with smaller numbers of participants from populations that may be difficult to recruit in large numbers. However, the convergence diagnostics from the current investigation make clear that this design may be in tension with the modeling demands of the approach leveraged here.

The Bayesian multilevel models used throughout this dissertation estimate target-level random effects that capture heterogeneity in cosine similarity across "chapters" of conversations within each target, and the Gaussian process hyperparameters require sufficient target-level variation to identify the temporal dynamics of interest. When the number of targets is small, these variance components are poorly identified because there are too few clusters at that level for the sampler to distinguish genuine between-target variability from sampling noise (Maas & Hox, 2005; McNeish & Stapleton, 2016). The Gaussian process component compounds the issue because both the amplitude and length-scale are estimated per speaker type, effectively doubling the number of hyperparameters that depend on target-level variation for identification.

The convergence patterns observed across the six studies in this investigation provide an empirical basis for estimating the minimum target-level sample size. The Negotiation study, which functioned well using the dyadic interaction paradigm, had 41 targets and showed stable convergence across all model specifications. If this set of analyses is viewed as a proxy for a power analysis, these patterns suggest that a minimum of approximately 40 targets would be necessary to reliably estimate the target-level random effects and GP hyperparameters that the current pipeline requires. This may be a meaningful contribution in its own right: prior to this investigation, there was no empirical basis in the inferential accuracy literature for specifying how many targets a standard stimulus study would need to support hierarchical Gaussian process modeling. The answer, roughly 40, is an order of magnitude larger than the typical standard stimulus sample (which often includes fewer than 10 targets). Future standard stimulus studies designed to test cue validity could aim to achieve this number of targets.

Three research questions emerge as particularly fruitful possible guides for future researchers in this context. First, do the paradigm-specific effects identified in Study 2

correspond to cue validity? That is, are the target questions that perceivers attend to early in chapters also reflective of what targets were thinking?

Second, does the null effect for speaker (target vs interactant) hold in the standard stimulus paradigm, or does the interviewer role that is often a part of this paradigm create genuine asymmetry in speech-thought similarity? The current investigation's finding that type of speaker doesn't seem to reliably predict utilization or validity in the dyadic interaction paradigm cannot be assumed to generalize to contexts where interactants and targets have fundamentally different conversational roles. In an interview, the interactant's primary function is to ask questions of the target and guide the conversation, not to express their own thoughts, and this asymmetry in conversational roles may produce asymmetry in cue validity not found in the more symmetric dyadic interaction paradigm.

Third, is the misalignment of cue utilization and cue validity for *suggestion* dialogue acts also present in the standard stimulus paradigm, or does the interview structure often used in the standard stimulus paradigm alter which dialogue acts are most semantically similar to target thoughts? In the standard stimulus paradigm, it is plausible that *questions* function as valid cues to target thought content, even though they did not in the dyadic interaction studies examined here (Huang et al., 2017). If so, the null finding about question dialogue acts reported in the current investigation could be reinterpreted as a paradigm-specific finding rather than a general feature of conversational cognition.

### *Tuning Dialogue Act Classification*

The DAMSL taxonomy (Core & Allen, 1997) may have been too coarse to capture the distinctions that matter for inferential accuracy. The seven-category collapsed structure was necessary to ensure sufficient observations for the Gaussian process models, and even then was

unable to converge for all dialogue acts in all studies, but it may have occluded meaningful variation within categories. The *question* category, in particular, collapsed open-ended and closed-ended questions into a single category, a distinction that may be theoretically relevant given that open-ended questions may elicit more disclosure and self-reflection than closed-ended ones. The *suggestion* dialogue act category, despite being the most valid cue identified in Study 3, had very low base rates (only 801 total observations across four studies), whereas the *inform* dialogue act category, by far the most common dialogue act in these conversations, may have been too heterogeneous, masking effects specific to the self-disclosing subset. There may be a "sweet spot" in specificity and representativeness that was over-generalized in the other dialogue acts.

These category-level limitations raise the question of whether the discursive findings reported here are artifacts of the DAMSL taxonomy or reflect genuine features of conversational cognition. This question can only be answered by replicating the analyses under alternative coding schemes. Several taxonomies could serve this purpose, each addressing a different limitation of the current approach.

An alternative is the Switchboard-DAMSL variant (Jurafsky et al., 1997), developed for the Switchboard corpus of telephone conversations. Its most relevant feature for the current investigation is a fine-grained question typology that distinguishes yes-no questions, wh-questions, open-ended questions, rhetorical questions, tag questions, and declarative yes-no questions. This directly addresses the concern that the null finding for Hypothesis 5 may be an artifact of aggregation: open-ended questions may genuinely elicit thought-reflective responses, but this effect could be washed out when combined with closed-ended or rhetorical questions that serve a different conversational function. The Switchboard-DAMSL also separates

statements that include opinions from statements that don't include opinions, providing a second route to decomposing the *inform* dialogue act category. A strategically collapsed version retaining the question subtypes and the opinion/non-opinion distinction while merging rare categories (approximately 12 tags) could maintain tractable category counts for the meta-analytic models.

Two other taxonomies address speech dimensions entirely absent from DAMSL. Bales' (1950) Interaction Process Analysis (IPA) introduces a socio-emotional dimension (positive and negative) alongside task-oriented acts, which may be relevant to inferential accuracy where the criterion includes targets' reported feelings as well as thoughts. Additionally, the "gives opinion" category in this taxonomy, defined to include evaluative speech expressing feelings or wishes, provides a principled way to distinguish substantive self-disclosure from factual information-giving. Another alternative taxonomy would be the DailyDialog taxonomy (Li et al., 2017), which introduces a commissive category (promises, offers, refusals) that has no direct DAMSL equivalent and may capture moments of cognitive commitment that correspond to target thought content, since when a speaker commits to a future action, they are expressing an intention likely to be reflected in what they are concurrently thinking.

The value of a cross-taxonomy comparison may be not in identifying the single best coding scheme but in using methodological variation as a source of information about the robustness of the findings. If the misalignment for *suggestion* dialogue acts replicates with Bales' "gives suggestion" category or DailyDialog's directive category, each of which draws different classification boundaries around similar dialogue acts, the finding may be more credibly a feature of conversational cognition than an artifact of any particular taxonomy. If the diffuse estimates for *question* dialogue acts resolve into a significant effect for open-ended questions

specifically under Switchboard-DAMSL but not under DAMSL's collapsed category, this could identify the aggregation as the source of the null findings rather than a genuine absence of effect. Findings that emerge consistently across taxonomies with different classification boundaries and different error structures may thus provide more robust evidence than any single taxonomy can offer (Grimmer et al., 2022), and the concern about misclassification bias is addressed more convincingly by cross-taxonomy consistency than by correction within a single scheme (TeBlunthuis et al., 2024).

Implementing such a multi-taxonomy approach has only recently become practical. Recent work has demonstrated that large language models achieve competitive accuracy on pragmatic classification tasks with minimal training examples (Gilardi et al., 2023), and the feasibility of automating fine-grained dialogue act annotation using large language models has been demonstrated with promising initial results (Su & Ye, 2025). This opens the possibility of annotating the full corpus of utterances using multiple taxonomies simultaneously, creating a multi-layered annotation structure in which each utterance carries parallel labels from several schemes. Such multi-label representations could then be leveraged using multi-task learning frameworks (e.g., Raffel et al., 2023) that model the relationships between taxonomies, potentially identifying patterns that no single coding scheme could detect. For example, a multi-task model could learn that utterances classified as *inform* dialogue acts under DAMSL but as "gives opinion" under Bales' IPA (the evaluative subset of *inform* dialogue acts) show different temporal dynamics than utterances that are classified as *inform* dialogue acts under DAMSL but as "gives orientation" acts under IPA (the factual subset). This approach moves beyond simply choosing between taxonomies and instead treats the choice of coding scheme as a source of information in its own right.

### *Open Conceptual Questions: What the Pipeline Reveals and What Remains*

Stepping back from the specific findings, the embedding-based approach raises three larger conceptual questions that earlier methods were not well positioned to ask because their available tools were too coarse to address them. The first concerns the level at which inferential accuracy operates: are perceivers tracking person-specific information about this target, or are they assembling inferences from a more generic conversational scaffold (e.g., recency, topic, dialogue norms) that would yield similar content for any plausible target in the same situation? The second concerns the direction of inference: this dissertation has treated the target's reported thought as the criterion against which a perceiver's inference is compared, but the null speaker effect suggests that the criterion itself is partly a product of the conversation, not an independent ground truth against which perceivers are measured. The third concerns the adequacy of semantic similarity as a proxy for accuracy. Similarity in an embedding space captures alignment in the distributional meaning of words, while inferential accuracy is a claim about correspondence between two mental representations (Hodges et al., 2015; Ickes et al., 1990) and a measure that could be inflated by topical convergence or shared stereotypes is not the same as a measure of mental-state correspondence, even where the two may be used interchangeably. Three speculative extensions, each pointing to a future direction, follow.

Did the target feel understood? Even when perceiver inferences are highly semantically similar to target thoughts, the target's subjective experience of being understood, often discussed in the close-relationships literature as perceived partner responsiveness (see Reis et al., 2017), is a separate construct that the current investigation did not measure. Content-level alignment does not guarantee the felt experience of being heard: a perceiver may accurately reproduce what a target was thinking while still missing the emotional emphasis, the reason the thought mattered,

or the aspect of the thought the target most hoped would be acknowledged. Future work pairing utterance-level similarity with target-reported felt understanding (e.g., via a post-interaction responsiveness measure) could ask whether the two constructs converge. If instead they diverge, it would suggest that the embedding-based approach captures one facet of what it means to be accurately inferred while leaving a distinct, experientially central facet unmeasured.

Is similarity driven by shared stereotypes rather than attention to the target? High semantic similarity between a perceiver's inference and a target's thought could reflect genuine attention to what the target said and thought, or it could reflect shared cultural scripts and stereotypes that cause perceiver and target to converge on similar content independently. Prior work in this tradition has established that stereotype use can facilitate accuracy (Hodges & Kezer, 2021; Lewis et al., 2012), but stereotype-driven convergence and attention-driven convergence are likely to produce similar embedding-level patterns. Future work could incorporate explicit measures of stereotype content and examine whether semantic similarity that aligns with stereotype content behaves differently from similarity that does not. A complementary, more conservative approach would be to compute a within-topic null distribution and evaluate observed similarity against that baseline (for use of baseline correction in inferential accuracy research, see Ickes et al., 1990, and Lewis et al., 2012).

**Analogue checks against topic convergence.** A closely related concern is that perceivers and targets may appear to align simply because they are both thinking about the same topic (i.e., the conversation's subject matter) rather than because any person-specific inference is taking place. Several extensions to this investigation could address this. First, a topic-only baseline condition would ask a sample of uninformed raters to generate inferences given only a brief description of the conversational topic (e.g., "a new mother describing her experience")

without access to the conversation itself; semantic similarity between these uninformed inferences and target thoughts would establish how much of the present findings can be attributed to topical convergence alone. Second, permutation analyses pairing each inference with a randomly selected, non-matched interactant or target's speech from the same study would provide an empirical "same topic, different person" baseline: if observed similarity does not exceed this shuffled baseline, the measure is indexing topic rather than person-specific inference. Third, embeddings generated from topic prompts alone could be compared to embeddings of actual perceiver inferences. None of these checks replaces the current analyses, but each would sharpen the inferential claims that can be drawn from them.

Taken together, these three questions and the extensions they suggest partially reframe the contribution of this dissertation. The embedding-based pipeline makes it possible to interrogate the temporal, structural, and discursive organization of inferential accuracy at a resolution not supported by earlier methods, revealing, for example, that suggestion dialogue acts carry more information about target thought content than any other discursive category examined, and that this information is not well reflected in what perceivers infer. That same resolution, however, exposes ambiguities that earlier methods could afford to ignore: similarity is not accuracy, topical convergence is not person-specific inference, and stereotype-driven content matching is not attention to the target. A fuller account of how perceivers make sense of what targets are thinking and feeling would be enhanced by pairing the utterance-level resolution introduced here with subjective, stereotype-controlled, and topic-controlled checks. This dissertation takes a substantive first step into that territory; the territory itself remains large.

## Conclusion

The three studies in this dissertation are, to my knowledge, the first to apply Brunswik's (1956) lens model to inferential accuracy at the resolution of individual utterances. Where previous lens model applications relied on researcher-defined cue sets (e.g., the behavioral cues in Bhowmik et al., 2025; the personality-relevant cues in Breil et al., 2021), the embedding-based approach taken here defines a cue space that encompasses the full conversational record. Every utterance in a recorded conversation becomes a potential cue whose similarity to perceiver inferences and target thoughts can be estimated, and the structural and discursive features of those utterances become the dimensions along which the cue space is organized.

The modeling demands imposed by this approach were considerable: Bayesian multilevel Gaussian process models, three-stage individual participant data meta-analytic pipelines, and pointwise posterior trajectory syntheses across heterogeneous studies. This complexity was necessary because the embedding-based approach generates a data structure that is too rich for simpler methods: each study contributes not a single effect size but a full nonlinear trajectory of semantic similarity over the course of a conversation, and these trajectories must be synthesized across studies differing in design, sample, and conversational structure.

What emerges from this complexity is a preliminary map of both sides of the lens for verbal information in inferential accuracy. For structural dimensions, the map suggests calibration: perceivers' inferences are most similar to recent speech, and recent speech is also most similar to what targets were thinking. Neither perceiver inferences nor target thoughts preferentially align with one speaker over the other, and the conversation, rather than any individual speaker's contributions, appears to be the primary carrier of information about target thought content. For the discursive dimensions, the map reveals a specific possible

miscalibration: *suggestions* are the dialogue act category most semantically similar to target thoughts (within dyadic interaction paradigm studies), yet perceiver inferences show no elevated similarity with *suggestions*.

Consider these findings in conjunction with the example from the Introduction: That millions of viewers watched "The Behavior Panel" (2021) dissect a single interview between Meghan Markle and Prince Harry (Winfrey, 2021) suggests that there is considerable interest in understanding what is going on "under the surface" of human conversations. However, this set of studies tells us that even just understanding what's *at* the surface (i.e., the words people say) offers an immense amount of complexity and, perhaps much to the chagrin of body language "influencers," seems particularly hard to pin down into bite-size conclusions. Instead, this dissertation offers a preliminary map, drawn at the resolution of individual utterances, of how those surface-level words may relate to the inferences formed by perceivers and the thoughts reported by targets. The embedding-based approach taken here reveals that the verbal cue space is structured in ways that are partially calibrated (perceivers attend to recent speech, and recent speech is indeed most reflective of target thoughts) and partially miscalibrated (the dialogue act category most similar to target thoughts is not the one most reflected in perceiver inferences). That this level of detail can be extracted from the conversational record of multiple studies analyzed in concert thus represents an exciting step towards further integrating novel natural language processing tools into inferential accuracy research, expanding the opportunities to further examine what is frequently described as the most predictive cue for inferential accuracy: verbal information.

# APPENDIX A

## STUDY 2 DIALOGUE ACT MODEL FIT STATISTICS

Beginning with Table A2, each table shows the estimates for all parameters within a single dialogue act. These retain the variable names from the original brms code, which are explained here and at the beginning of Appendix B.

**speaker_typetarget** and **speaker_typepartner**: Mean cosine similarity for target or interactant speech. Because the model drops the intercept (the `0 +` syntax), this is the mean, not a difference from some reference group.

**I(percent_of_chapter_at_utterance_start)^2**: This is a labeling artifact from the HSGP approximation. The actual variable underneath is the normalized position of each utterance in its chapter (0 = start, 1 = end) — i.e., the time axis that the Gaussian process is smoothing over.

**sigma**: Residual standard deviation; whatever observation-level noise is left after the fixed effects, GP, and random effects.

**sd(Intercept)** — shows up three times per study because there are three nested grouping factors. *Perceiver* (or target, in Appendix B): How much individual perceivers differ from each other in baseline cosine similarity. *Target*: Between-target variability in the sample. *Chapter*: How much cosine similarity bounces around from chapter to chapter within the same target.

***Table A1:***

*Convergence Diagnostics for Study 2 Perceiver Dialogue Act Models*

| Hypothesis | Dialogue Act | Study | Max Rhat | Min ESS | Divergent |
|---|---|---|---|---|---|
| H5 | Acknowledgment | STEM | 1.002 | 2,123 | 173 |
| | | Middle Eastern Men | 1.002 | 655 | 4 |
| | | Negotiation | 1.015 | 332 | 482 |
| | | Round Robin (Video Mediated) | 1.239 | 13 | 400 |
| | | Round Robin (In Person) | 1.002 | 1,544 | 19 |
| | Agreement | New Mothers | 1.001 | 7,093 | 0 |
| | | STEM | 1.008 | 555 | 17 |
| | | Middle Eastern Men | 1.159 | 18 | 237 |
| | | Negotiation | 1.001 | 4,517 | 2 |
| | | Round Robin (Video Mediated) | 1.003 | 1,348 | 23 |
| | | Round Robin (In Person) | 1.012 | 229 | 3 |
| | Disagreement | STEM | 1.014 | 716 | 9 |
| | | Negotiation | 1.002 | 2,473 | 0 |
| | | Round Robin (Video Mediated) | 1.002 | 1,541 | 6 |
| | | Round Robin (In Person) | 1.011 | 517 | 9 |
| | Functional | STEM | 1.008 | 736 | 2 |
| | | Negotiation | 1.000 | 2,900 | 0 |
| | | Round Robin (Video Mediated) | 1.003 | 1,925 | 3 |
| | | Round Robin (In Person) | 1.005 | 829 | 0 |
| | Inform | New Mothers | 1.001 | 4,312 | 0 |
| | | STEM | 1.002 | 560 | 4 |
| | | Negotiation | 1.002 | 2,100 | 0 |
| | | Round Robin (Video Mediated) | 1.001 | 2,034 | 23 |
| | | Round Robin (In Person) | 1.011 | 317 | 3 |
| | Question | New Mothers | 1.001 | 2,942 | 0 |

| Hypothesis | Dialogue Act | Study | Max Rhat | Min ESS | Divergent |
|---|---|---|---|---|---|
| | | STEM | 1.003 | 829 | 0 |
| | | Middle Eastern Men | 1.010 | 549 | 0 |
| | | Negotiation | 1.001 | 3,435 | 2 |
| | | Round Robin (Video Mediated) | 1.002 | 2,099 | 6 |
| | | Round Robin (In Person) | 1.006 | 434 | 3 |
| | Suggestion | STEM | 1.002 | 1,183 | 0 |
| | | Negotiation | 1.002 | 2,663 | 92 |
| | | Round Robin (Video Mediated) | 1.002 | 632 | 1 |
| | | Round Robin (In Person) | 1.020 | 602 | 0 |
| H6 | Acknowledgment | STEM | 1.001 | 8,168 | 0 |
| | | Negotiation | 1.504 | 24 | 2897 |
| | | Round Robin (Video Mediated) | 1.001 | 5,861 | 139 |
| | | Round Robin (In Person) | 1.003 | 1,202 | 0 |
| | Agreement | New Mothers | 1.002 | 1,623 | 0 |
| | | STEM | 1.002 | 1,709 | 10 |
| | | Middle Eastern Men | 1.000 | 2,852 | 0 |
| | | Negotiation | 1.001 | 3,294 | 7 |
| | | Round Robin (Video Mediated) | 1.002 | 2,285 | 35 |
| | | Round Robin (In Person) | 1.002 | 1,372 | 19 |
| | Disagreement | New Mothers | 1.005 | 1,681 | 0 |
| | | STEM | 1.003 | 618 | 0 |
| | | Negotiation | 1.002 | 2,644 | 0 |
| | | Round Robin (Video Mediated) | 1.003 | 1,262 | 92 |
| | | Round Robin (In Person) | 1.003 | 1,039 | 22 |
| | Functional | STEM | 1.003 | 1,810 | 0 |
| | | Negotiation | 1.000 | 5,938 | 0 |
| | | Round Robin (Video Mediated) | 1.002 | 1,747 | 7 |

| Hypothesis | Dialogue Act | Study | Max Rhat | Min ESS | Divergent |
|---|---|---|---|---|---|
| | | Round Robin (In Person) | 1.002 | 1,415 | 3 |
| | Inform | New Mothers | 1.001 | 3,081 | 0 |
| | | STEM | 1.007 | 610 | 12 |
| | | Negotiation | 1.001 | 3,514 | 2 |
| | | Round Robin (Video Mediated) | 1.001 | 1,498 | 22 |
| | | Round Robin (In Person) | 1.028 | 285 | 25 |
| | Question | New Mothers | 1.000 | 4,758 | 0 |
| | | STEM | 1.006 | 1,041 | 1 |
| | | Negotiation | 1.002 | 2,549 | 0 |
| | | Round Robin (Video Mediated) | 1.002 | 2,334 | 9 |
| | | Round Robin (In Person) | 1.008 | 675 | 4 |
| | Suggestion | STEM | 1.001 | 2,984 | 0 |
| | | Negotiation | 1.000 | 6,773 | 0 |
| | | Round Robin (Video Mediated) | 1.001 | 3,933 | 0 |
| | | Round Robin (In Person) | 1.005 | 1,295 | 0 |

***Table A2:***

*H5: Acknowledgment Parameter Estimates*

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| STEM | speaker_typetarget | Fixed | 0.035 | 0.049 | -0.060 | 0.134 | 1.002 | 2,388.383 |
| | speaker_typepartner | Fixed | 0.143 | 0.053 | 0.042 | 0.251 | 1.001 | 6,413.504 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.240 | 0.078 | 0.083 | 0.398 | 1.002 | 2,122.991 |
| | sigma | Residual | 0.043 | 0.031 | 0.003 | 0.113 | 1.014 | 266.677 |
| | sd(Intercept) | Random | 0.050 | 0.035 | 0.002 | 0.124 | 1.001 | 3,144.871 |
| | sd(Intercept) | Random | 0.051 | 0.034 | 0.002 | 0.121 | 1.002 | 3,272.882 |
| | sd(Intercept) | Random | 0.041 | 0.031 | 0.002 | 0.112 | 1.001 | 3,110.392 |
| Middle Eastern Men | speaker_typetarget | Fixed | 0.116 | 0.047 | 0.019 | 0.213 | 1.000 | 6,711.706 |
| | speaker_typepartner | Fixed | 0.089 | 0.044 | 0.002 | 0.179 | 1.000 | 6,403.471 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.056 | 0.045 | -0.029 | 0.156 | 1.000 | 7,166.497 |
| | sigma | Residual | 0.031 | 0.002 | 0.028 | 0.034 | 1.000 | 3,785.036 |
| | sd(Intercept) | Random | 0.010 | 0.006 | 0.000 | 0.023 | 1.002 | 654.762 |
| | sd(Intercept) | Random | 0.059 | 0.043 | 0.009 | 0.173 | 1.001 | 5,655.099 |
| | sd(Intercept) | Random | 0.050 | 0.003 | 0.044 | 0.056 | 1.001 | 2,316.147 |
| Negotiation | speaker_typetarget | Fixed | 0.159 | 0.091 | -0.024 | 0.333 | 1.004 | 1,293.071 |
| | speaker_typepartner | Fixed | 0.105 | 0.083 | -0.058 | 0.271 | 1.002 | 4,329.219 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.076 | 0.105 | -0.120 | 0.297 | 1.000 | 6,275.181 |
| | sigma | Residual | 0.052 | 0.044 | 0.001 | 0.164 | 1.034 | 149.097 |
| | sd(Intercept) | Random | 0.084 | 0.062 | 0.003 | 0.226 | 1.005 | 2,193.584 |
| | sd(Intercept) | Random | 0.082 | 0.063 | 0.000 | 0.226 | 1.015 | 332.368 |
| | sd(Intercept) | Random | 0.062 | 0.053 | 0.001 | 0.194 | 1.011 | 444.940 |
| Round Robin (Video Mediated) | speaker_typetarget | Fixed | 0.061 | 0.024 | 0.027 | 0.116 | 1.239 | 12.879 |
| | speaker_typepartner | Fixed | 0.113 | 0.023 | 0.068 | 0.157 | 1.179 | 15.565 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.009 | 0.032 | -0.058 | 0.086 | 1.169 | 86.129 |
| | sigma | Residual | 0.007 | 0.010 | 0.000 | 0.037 | 1.179 | 16.647 |
| | sd(Intercept) | Random | 0.046 | 0.023 | 0.003 | 0.083 | 1.099 | 37.488 |
| | sd(Intercept) | Random | 0.041 | 0.019 | 0.003 | 0.079 | 1.163 | 102.656 |
| | sd(Intercept) | Random | 0.034 | 0.023 | 0.003 | 0.082 | 1.185 | 15.785 |
| | speaker_typetarget | Fixed | 0.127 | 0.017 | 0.096 | 0.154 | 1.001 | 4,601.342 |

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| Round Robin (In Person) | speaker_typepartner | Fixed | 0.114 | 0.015 | 0.086 | 0.142 | 1.000 | 5,336.918 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.011 | 0.020 | -0.028 | 0.051 | 1.000 | 5,546.961 |
| | sigma | Residual | 0.031 | 0.007 | 0.020 | 0.046 | 1.001 | 1,514.472 |
| | sd(Intercept) | Random | 0.013 | 0.009 | 0.001 | 0.033 | 1.000 | 3,070.606 |
| | sd(Intercept) | Random | 0.014 | 0.009 | 0.001 | 0.034 | 1.002 | 2,864.857 |
| | sd(Intercept) | Random | 0.032 | 0.010 | 0.008 | 0.048 | 1.001 | 1,543.584 |

***Table A3:***

*H5: Agreement Parameter Estimates*

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| New Mothers | speaker_typetarget | Fixed | 0.100 | 0.064 | -0.027 | 0.227 | 1.000 | 8,141.315 |
| | speaker_typepartner | Fixed | 0.134 | 0.063 | 0.011 | 0.258 | 1.000 | 7,824.789 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.113 | 0.096 | -0.069 | 0.307 | 1.000 | 8,042.167 |
| | sigma | Residual | 0.090 | 0.002 | 0.086 | 0.093 | 1.000 | 7,329.310 |
| | sd(Intercept) | Random | 0.003 | 0.003 | 0.000 | 0.009 | 1.000 | 7,092.823 |
| | sd(Intercept) | Random | 0.069 | 0.027 | 0.035 | 0.138 | 1.000 | 7,174.094 |
| | sd(Intercept) | Random | 0.004 | 0.003 | 0.000 | 0.010 | 1.001 | 7,220.189 |
| STEM | speaker_typetarget | Fixed | 0.117 | 0.013 | 0.081 | 0.134 | 1.000 | 4,898.588 |
| | speaker_typepartner | Fixed | 0.133 | 0.007 | 0.122 | 0.144 | 1.000 | 5,430.791 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.035 | 0.008 | 0.019 | 0.050 | 1.000 | 8,175.838 |
| | sigma | Residual | 0.110 | 0.001 | 0.107 | 0.112 | 1.000 | 6,515.884 |
| | sd(Intercept) | Random | 0.013 | 0.007 | 0.001 | 0.025 | 1.008 | 585.745 |
| | sd(Intercept) | Random | 0.014 | 0.007 | 0.001 | 0.025 | 1.008 | 555.214 |
| | sd(Intercept) | Random | 0.045 | 0.003 | 0.040 | 0.050 | 1.000 | 4,885.367 |
| Middle Eastern Men | speaker_typetarget | Fixed | 0.066 | 0.076 | -0.114 | 0.236 | 1.134 | 1,851.177 |
| | speaker_typepartner | Fixed | 0.098 | 0.077 | -0.053 | 0.255 | 1.025 | 254.814 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.034 | 0.103 | -0.223 | 0.161 | 1.143 | 18.942 |
| | sigma | Residual | 0.079 | 0.001 | 0.077 | 0.080 | 1.003 | 2,062.975 |
| | sd(Intercept) | Random | 0.002 | 0.001 | 0.000 | 0.005 | 1.000 | 1,049.841 |
| | sd(Intercept) | Random | 0.049 | 0.018 | 0.025 | 0.098 | 1.013 | 847.064 |
| | sd(Intercept) | Random | 0.036 | 0.002 | 0.032 | 0.040 | 1.159 | 17.893 |
| Negotiation | speaker_typetarget | Fixed | 0.116 | 0.010 | 0.096 | 0.133 | 1.001 | 6,102.196 |
| | speaker_typepartner | Fixed | 0.107 | 0.010 | 0.089 | 0.125 | 1.000 | 7,027.873 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.051 | 0.014 | 0.024 | 0.079 | 1.000 | 7,619.420 |
| | sigma | Residual | 0.090 | 0.002 | 0.086 | 0.095 | 1.000 | 7,202.683 |
| | sd(Intercept) | Random | 0.007 | 0.005 | 0.000 | 0.018 | 1.000 | 4,779.781 |
| | sd(Intercept) | Random | 0.007 | 0.005 | 0.000 | 0.018 | 1.000 | 4,517.384 |
| | sd(Intercept) | Random | 0.039 | 0.005 | 0.030 | 0.048 | 1.000 | 5,411.728 |
| | speaker_typetarget | Fixed | 0.125 | 0.004 | 0.116 | 0.131 | 1.000 | 6,626.608 |

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| Round Robin (Video Mediated) | speaker_typepartner | Fixed | 0.125 | 0.005 | 0.116 | 0.134 | 1.001 | 6,987.713 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.001 | 0.005 | -0.007 | 0.011 | 1.001 | 5,795.254 |
| | sigma | Residual | 0.055 | 0.000 | 0.055 | 0.056 | 1.000 | 7,157.952 |
| | sd(Intercept) | Random | 0.015 | 0.001 | 0.012 | 0.018 | 1.001 | 4,643.839 |
| | sd(Intercept) | Random | 0.005 | 0.002 | 0.001 | 0.010 | 1.003 | 1,347.513 |
| | sd(Intercept) | Random | 0.034 | 0.001 | 0.033 | 0.036 | 1.001 | 4,613.032 |
| Round Robin (In Person) | speaker_typetarget | Fixed | 0.118 | 0.003 | 0.114 | 0.124 | 1.000 | 6,169.846 |
| | speaker_typepartner | Fixed | 0.117 | 0.002 | 0.113 | 0.121 | 1.001 | 5,894.093 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.001 | 0.003 | -0.004 | 0.006 | 1.000 | 8,265.546 |
| | sigma | Residual | 0.057 | 0.000 | 0.056 | 0.058 | 1.000 | 7,390.621 |
| | sd(Intercept) | Random | 0.009 | 0.004 | 0.001 | 0.016 | 1.010 | 229.397 |
| | sd(Intercept) | Random | 0.009 | 0.004 | 0.000 | 0.015 | 1.012 | 232.770 |
| | sd(Intercept) | Random | 0.036 | 0.001 | 0.034 | 0.038 | 1.002 | 3,829.418 |

***Table A4:***

*H5: Disagreement Parameter Estimates*

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| STEM | speaker_typetarget | Fixed | 0.172 | 0.015 | 0.136 | 0.194 | 1.000 | 5,781.379 |
| | speaker_typepartner | Fixed | 0.170 | 0.012 | 0.147 | 0.189 | 1.000 | 6,644.432 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.028 | 0.013 | 0.005 | 0.054 | 1.000 | 7,351.644 |
| | sigma | Residual | 0.130 | 0.002 | 0.126 | 0.135 | 1.000 | 5,852.637 |
| | sd(Intercept) | Random | 0.020 | 0.010 | 0.001 | 0.036 | 1.014 | 733.058 |
| | sd(Intercept) | Random | 0.021 | 0.010 | 0.001 | 0.037 | 1.013 | 716.259 |
| | sd(Intercept) | Random | 0.061 | 0.004 | 0.052 | 0.069 | 1.000 | 4,544.190 |
| Negotiation | speaker_typetarget | Fixed | 0.123 | 0.022 | 0.080 | 0.167 | 1.000 | 7,615.196 |
| | speaker_typepartner | Fixed | 0.125 | 0.023 | 0.081 | 0.170 | 1.000 | 8,007.596 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.122 | 0.033 | 0.058 | 0.188 | 1.000 | 8,339.145 |
| | sigma | Residual | 0.122 | 0.007 | 0.108 | 0.135 | 1.001 | 4,934.192 |
| | sd(Intercept) | Random | 0.039 | 0.023 | 0.002 | 0.084 | 1.001 | 2,472.746 |
| | sd(Intercept) | Random | 0.039 | 0.023 | 0.002 | 0.084 | 1.001 | 2,588.130 |
| | sd(Intercept) | Random | 0.031 | 0.020 | 0.001 | 0.071 | 1.002 | 2,692.655 |
| Round Robin (Video Mediated) | speaker_typetarget | Fixed | 0.114 | 0.011 | 0.089 | 0.134 | 1.000 | 5,274.465 |
| | speaker_typepartner | Fixed | 0.116 | 0.007 | 0.107 | 0.130 | 1.002 | 4,727.444 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.002 | 0.010 | -0.022 | 0.019 | 1.001 | 4,731.174 |
| | sigma | Residual | 0.073 | 0.001 | 0.071 | 0.075 | 1.000 | 5,855.132 |
| | sd(Intercept) | Random | 0.013 | 0.003 | 0.007 | 0.017 | 1.002 | 2,578.529 |
| | sd(Intercept) | Random | 0.007 | 0.003 | 0.001 | 0.013 | 1.001 | 1,541.266 |
| | sd(Intercept) | Random | 0.034 | 0.002 | 0.030 | 0.038 | 1.001 | 3,477.277 |
| Round Robin (In Person) | speaker_typetarget | Fixed | 0.109 | 0.006 | 0.099 | 0.119 | 1.000 | 6,314.160 |
| | speaker_typepartner | Fixed | 0.097 | 0.021 | 0.047 | 0.129 | 1.000 | 6,185.746 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.001 | 0.010 | -0.017 | 0.021 | 1.000 | 6,937.441 |
| | sigma | Residual | 0.073 | 0.001 | 0.071 | 0.075 | 1.000 | 6,173.257 |
| | sd(Intercept) | Random | 0.010 | 0.005 | 0.001 | 0.019 | 1.011 | 517.496 |
| | sd(Intercept) | Random | 0.010 | 0.005 | 0.001 | 0.018 | 1.011 | 554.062 |
| | sd(Intercept) | Random | 0.033 | 0.002 | 0.029 | 0.037 | 1.001 | 3,310.971 |

***Table A5:***

*H5: Functional Parameter Estimates*

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| STEM | speaker_typetarget | Fixed | 0.187 | 0.019 | 0.142 | 0.217 | 1.001 | 5,658.904 |
| | speaker_typepartner | Fixed | 0.171 | 0.014 | 0.143 | 0.196 | 1.000 | 7,204.747 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.014 | 0.019 | -0.020 | 0.053 | 1.000 | 6,483.268 |
| | sigma | Residual | 0.127 | 0.003 | 0.121 | 0.133 | 1.000 | 4,267.880 |
| | sd(Intercept) | Random | 0.023 | 0.012 | 0.001 | 0.044 | 1.008 | 735.806 |
| | sd(Intercept) | Random | 0.023 | 0.013 | 0.001 | 0.044 | 1.005 | 737.416 |
| | sd(Intercept) | Random | 0.073 | 0.006 | 0.061 | 0.085 | 1.000 | 3,311.856 |
| Negotiation | speaker_typetarget | Fixed | 0.145 | 0.028 | 0.091 | 0.199 | 1.000 | 7,735.794 |
| | speaker_typepartner | Fixed | 0.160 | 0.028 | 0.107 | 0.215 | 1.000 | 7,494.889 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.100 | 0.045 | 0.012 | 0.190 | 1.000 | 8,527.715 |
| | sigma | Residual | 0.136 | 0.011 | 0.115 | 0.158 | 1.000 | 4,187.123 |
| | sd(Intercept) | Random | 0.025 | 0.018 | 0.001 | 0.066 | 1.000 | 5,534.290 |
| | sd(Intercept) | Random | 0.025 | 0.018 | 0.001 | 0.067 | 1.000 | 5,302.609 |
| | sd(Intercept) | Random | 0.042 | 0.025 | 0.002 | 0.092 | 1.000 | 2,900.242 |
| Round Robin (Video Mediated) | speaker_typetarget | Fixed | 0.128 | 0.011 | 0.106 | 0.145 | 1.001 | 6,714.022 |
| | speaker_typepartner | Fixed | 0.129 | 0.008 | 0.114 | 0.144 | 1.001 | 7,750.946 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | -0.020 | 0.015 | -0.049 | 0.011 | 1.001 | 7,710.748 |
| | sigma | Residual | 0.079 | 0.003 | 0.074 | 0.085 | 1.001 | 3,375.993 |
| | sd(Intercept) | Random | 0.024 | 0.006 | 0.010 | 0.034 | 1.000 | 2,085.752 |
| | sd(Intercept) | Random | 0.021 | 0.007 | 0.006 | 0.033 | 1.003 | 1,925.343 |
| | sd(Intercept) | Random | 0.045 | 0.006 | 0.034 | 0.056 | 1.002 | 1,975.653 |
| Round Robin (In Person) | speaker_typetarget | Fixed | 0.095 | 0.006 | 0.084 | 0.107 | 1.000 | 7,057.776 |
| | speaker_typepartner | Fixed | 0.097 | 0.007 | 0.086 | 0.109 | 1.000 | 6,447.937 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.002 | 0.009 | -0.017 | 0.020 | 1.000 | 7,956.548 |
| | sigma | Residual | 0.071 | 0.002 | 0.067 | 0.075 | 1.001 | 4,442.616 |
| | sd(Intercept) | Random | 0.012 | 0.007 | 0.001 | 0.025 | 1.005 | 828.533 |
| | sd(Intercept) | Random | 0.012 | 0.007 | 0.001 | 0.025 | 1.004 | 893.903 |
| | sd(Intercept) | Random | 0.044 | 0.004 | 0.036 | 0.052 | 1.001 | 2,590.939 |

***Table A6:***

*H5: Inform Parameter Estimates*

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| New Mothers | speaker_typetarget | Fixed | 0.141 | 0.022 | 0.099 | 0.186 | 1.000 | 4,312.346 |
| | speaker_typepartner | Fixed | 0.204 | 0.068 | 0.055 | 0.334 | 1.000 | 7,647.370 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.042 | 0.029 | -0.021 | 0.092 | 1.000 | 6,480.090 |
| | sigma | Residual | 0.101 | 0.001 | 0.099 | 0.103 | 1.000 | 7,586.763 |
| | sd(Intercept) | Random | 0.018 | 0.002 | 0.014 | 0.023 | 1.001 | 5,464.619 |
| | sd(Intercept) | Random | 0.040 | 0.015 | 0.021 | 0.077 | 1.000 | 5,820.326 |
| | sd(Intercept) | Random | 0.033 | 0.002 | 0.029 | 0.036 | 1.000 | 5,205.022 |
| STEM | speaker_typetarget | Fixed | 0.161 | 0.015 | 0.121 | 0.182 | 1.000 | 6,221.762 |
| | speaker_typepartner | Fixed | 0.163 | 0.007 | 0.151 | 0.176 | 1.000 | 6,918.082 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.022 | 0.008 | 0.006 | 0.038 | 1.001 | 8,155.326 |
| | sigma | Residual | 0.124 | 0.001 | 0.123 | 0.126 | 1.000 | 7,455.275 |
| | sd(Intercept) | Random | 0.019 | 0.010 | 0.001 | 0.033 | 1.001 | 560.217 |
| | sd(Intercept) | Random | 0.020 | 0.010 | 0.001 | 0.034 | 1.002 | 561.119 |
| | sd(Intercept) | Random | 0.055 | 0.002 | 0.051 | 0.059 | 1.000 | 5,532.814 |
| Negotiation | speaker_typetarget | Fixed | 0.143 | 0.015 | 0.114 | 0.172 | 1.000 | 8,200.451 |
| | speaker_typepartner | Fixed | 0.130 | 0.012 | 0.107 | 0.152 | 1.000 | 8,903.888 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.119 | 0.017 | 0.086 | 0.155 | 1.000 | 9,010.789 |
| | sigma | Residual | 0.134 | 0.003 | 0.128 | 0.139 | 1.000 | 6,240.697 |
| | sd(Intercept) | Random | 0.022 | 0.012 | 0.001 | 0.044 | 1.002 | 2,135.269 |
| | sd(Intercept) | Random | 0.022 | 0.012 | 0.001 | 0.044 | 1.001 | 2,099.727 |
| | sd(Intercept) | Random | 0.038 | 0.006 | 0.026 | 0.050 | 1.000 | 5,075.068 |
| Round Robin (Video Mediated) | speaker_typetarget | Fixed | 0.118 | 0.003 | 0.112 | 0.125 | 1.001 | 6,362.284 |
| | speaker_typepartner | Fixed | 0.118 | 0.003 | 0.113 | 0.125 | 1.001 | 5,433.768 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | -0.003 | 0.004 | -0.010 | 0.005 | 1.001 | 5,340.865 |
| | sigma | Residual | 0.076 | 0.000 | 0.075 | 0.076 | 1.001 | 6,900.186 |
| | sd(Intercept) | Random | 0.015 | 0.001 | 0.012 | 0.017 | 1.000 | 4,372.942 |
| | sd(Intercept) | Random | 0.008 | 0.002 | 0.004 | 0.012 | 1.000 | 2,034.410 |
| | sd(Intercept) | Random | 0.034 | 0.001 | 0.032 | 0.036 | 1.000 | 4,765.818 |
| | speaker_typetarget | Fixed | 0.110 | 0.005 | 0.100 | 0.118 | 1.000 | 4,700.653 |

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| Round Robin (In Person) | speaker_typepartner | Fixed | 0.101 | 0.016 | 0.062 | 0.126 | 1.000 | 5,422.135 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.004 | 0.006 | -0.005 | 0.020 | 1.001 | 5,135.025 |
| | sigma | Residual | 0.076 | 0.000 | 0.076 | 0.077 | 1.000 | 7,615.739 |
| | sd(Intercept) | Random | 0.007 | 0.003 | 0.000 | 0.012 | 1.006 | 353.451 |
| | sd(Intercept) | Random | 0.007 | 0.004 | 0.000 | 0.013 | 1.011 | 316.998 |
| | sd(Intercept) | Random | 0.035 | 0.001 | 0.034 | 0.037 | 1.001 | 3,749.611 |

***Table A7:***

*H5: Question Parameter Estimates*

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| New Mothers | speaker_typetarget | Fixed | 0.130 | 0.051 | 0.026 | 0.232 | 1.000 | 6,585.013 |
| | speaker_typepartner | Fixed | 0.161 | 0.043 | 0.075 | 0.244 | 1.001 | 6,086.306 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.112 | 0.060 | -0.002 | 0.234 | 1.000 | 6,905.839 |
| | sigma | Residual | 0.090 | 0.002 | 0.086 | 0.095 | 1.002 | 3,384.436 |
| | sd(Intercept) | Random | 0.036 | 0.004 | 0.028 | 0.045 | 1.000 | 4,769.845 |
| | sd(Intercept) | Random | 0.056 | 0.020 | 0.030 | 0.107 | 1.000 | 5,811.817 |
| | sd(Intercept) | Random | 0.061 | 0.004 | 0.052 | 0.069 | 1.001 | 2,941.559 |
| STEM | speaker_typetarget | Fixed | 0.209 | 0.011 | 0.189 | 0.230 | 1.000 | 7,733.458 |
| | speaker_typepartner | Fixed | 0.206 | 0.011 | 0.185 | 0.226 | 1.000 | 7,092.780 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | -0.019 | 0.014 | -0.046 | 0.009 | 1.001 | 7,752.330 |
| | sigma | Residual | 0.135 | 0.002 | 0.130 | 0.140 | 1.001 | 5,897.618 |
| | sd(Intercept) | Random | 0.020 | 0.011 | 0.001 | 0.038 | 1.003 | 829.306 |
| | sd(Intercept) | Random | 0.020 | 0.011 | 0.001 | 0.038 | 1.002 | 921.063 |
| | sd(Intercept) | Random | 0.076 | 0.005 | 0.066 | 0.085 | 1.001 | 4,762.608 |
| Middle Eastern Men | speaker_typetarget | Fixed | 0.172 | 0.142 | -0.106 | 0.448 | 1.002 | 1,229.936 |
| | speaker_typepartner | Fixed | 0.038 | 0.125 | -0.207 | 0.280 | 1.003 | 1,354.209 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | -0.000 | 0.140 | -0.266 | 0.274 | 1.003 | 1,752.644 |
| | sigma | Residual | 0.104 | 0.001 | 0.103 | 0.106 | 1.002 | 1,415.447 |
| | sd(Intercept) | Random | 0.023 | 0.002 | 0.019 | 0.027 | 1.005 | 708.668 |
| | sd(Intercept) | Random | 0.049 | 0.017 | 0.027 | 0.093 | 1.010 | 549.153 |
| | sd(Intercept) | Random | 0.047 | 0.002 | 0.044 | 0.050 | 1.001 | 806.089 |
| Negotiation | speaker_typetarget | Fixed | 0.140 | 0.025 | 0.090 | 0.189 | 1.000 | 8,262.019 |
| | speaker_typepartner | Fixed | 0.134 | 0.024 | 0.086 | 0.177 | 1.000 | 7,544.734 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.081 | 0.037 | 0.004 | 0.151 | 1.000 | 8,332.792 |
| | sigma | Residual | 0.134 | 0.006 | 0.123 | 0.146 | 1.001 | 5,436.988 |
| | sd(Intercept) | Random | 0.017 | 0.012 | 0.001 | 0.045 | 1.000 | 4,799.374 |
| | sd(Intercept) | Random | 0.018 | 0.012 | 0.001 | 0.045 | 1.000 | 4,668.362 |
| | sd(Intercept) | Random | 0.041 | 0.015 | 0.007 | 0.067 | 1.001 | 3,435.490 |
| | speaker_typetarget | Fixed | 0.117 | 0.005 | 0.108 | 0.126 | 1.000 | 6,966.442 |

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| Round Robin (Video Mediated) | speaker_typepartner | Fixed | 0.115 | 0.005 | 0.107 | 0.123 | 1.001 | 7,259.902 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.009 | 0.006 | -0.002 | 0.022 | 1.000 | 7,845.864 |
| | sigma | Residual | 0.082 | 0.001 | 0.080 | 0.084 | 1.001 | 6,135.886 |
| | sd(Intercept) | Random | 0.020 | 0.002 | 0.016 | 0.024 | 1.000 | 5,366.314 |
| | sd(Intercept) | Random | 0.005 | 0.003 | 0.000 | 0.012 | 1.002 | 2,098.931 |
| | sd(Intercept) | Random | 0.035 | 0.002 | 0.031 | 0.039 | 1.002 | 4,026.968 |
| Round Robin (In Person) | speaker_typetarget | Fixed | 0.107 | 0.004 | 0.101 | 0.113 | 1.000 | 6,251.169 |
| | speaker_typepartner | Fixed | 0.104 | 0.004 | 0.097 | 0.111 | 1.001 | 5,987.730 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.005 | 0.005 | -0.004 | 0.014 | 1.000 | 7,665.051 |
| | sigma | Residual | 0.077 | 0.001 | 0.076 | 0.079 | 1.000 | 6,139.808 |
| | sd(Intercept) | Random | 0.010 | 0.005 | 0.001 | 0.018 | 1.006 | 434.206 |
| | sd(Intercept) | Random | 0.010 | 0.005 | 0.000 | 0.018 | 1.006 | 461.179 |
| | sd(Intercept) | Random | 0.033 | 0.002 | 0.029 | 0.036 | 1.001 | 4,107.968 |

***Table A8:***

*H5: Suggestion Parameter Estimates*

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| STEM | speaker_typetarget | Fixed | 0.237 | 0.025 | 0.187 | 0.287 | 1.000 | 7,215.910 |
| | speaker_typepartner | Fixed | 0.193 | 0.026 | 0.142 | 0.242 | 1.001 | 7,216.225 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | -0.028 | 0.034 | -0.096 | 0.041 | 1.000 | 7,166.170 |
| | sigma | Residual | 0.097 | 0.011 | 0.077 | 0.122 | 1.001 | 1,328.503 |
| | sd(Intercept) | Random | 0.017 | 0.012 | 0.001 | 0.045 | 1.002 | 3,510.466 |
| | sd(Intercept) | Random | 0.017 | 0.012 | 0.001 | 0.045 | 1.001 | 3,643.193 |
| | sd(Intercept) | Random | 0.073 | 0.019 | 0.023 | 0.102 | 1.002 | 1,183.075 |
| Negotiation | speaker_typetarget | Fixed | 0.176 | 0.070 | 0.032 | 0.312 | 1.001 | 4,005.842 |
| | speaker_typepartner | Fixed | 0.117 | 0.068 | -0.019 | 0.250 | 1.001 | 6,132.790 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.114 | 0.087 | -0.058 | 0.287 | 1.000 | 8,098.247 |
| | sigma | Residual | 0.073 | 0.039 | 0.011 | 0.160 | 1.004 | 1,141.911 |
| | sd(Intercept) | Random | 0.054 | 0.044 | 0.002 | 0.163 | 1.000 | 5,749.455 |
| | sd(Intercept) | Random | 0.054 | 0.045 | 0.002 | 0.168 | 1.002 | 2,663.040 |
| | sd(Intercept) | Random | 0.062 | 0.043 | 0.003 | 0.158 | 1.000 | 4,602.825 |
| Round Robin (Video Mediated) | speaker_typetarget | Fixed | 0.134 | 0.020 | 0.089 | 0.168 | 1.000 | 6,124.955 |
| | speaker_typepartner | Fixed | 0.143 | 0.023 | 0.105 | 0.192 | 1.000 | 5,234.709 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.001 | 0.038 | -0.070 | 0.080 | 1.000 | 4,212.552 |
| | sigma | Residual | 0.069 | 0.010 | 0.052 | 0.090 | 1.003 | 697.184 |
| | sd(Intercept) | Random | 0.013 | 0.009 | 0.001 | 0.034 | 1.000 | 2,845.884 |
| | sd(Intercept) | Random | 0.013 | 0.009 | 0.001 | 0.035 | 1.001 | 2,659.532 |
| | sd(Intercept) | Random | 0.054 | 0.017 | 0.009 | 0.077 | 1.002 | 631.579 |
| Round Robin (In Person) | speaker_typetarget | Fixed | 0.120 | 0.015 | 0.092 | 0.149 | 1.001 | 7,251.209 |
| | speaker_typepartner | Fixed | 0.134 | 0.016 | 0.104 | 0.163 | 1.000 | 7,569.079 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.030 | 0.026 | -0.021 | 0.082 | 1.000 | 8,608.886 |
| | sigma | Residual | 0.060 | 0.011 | 0.043 | 0.084 | 1.016 | 707.626 |
| | sd(Intercept) | Random | 0.026 | 0.016 | 0.001 | 0.057 | 1.001 | 1,425.931 |
| | sd(Intercept) | Random | 0.025 | 0.016 | 0.001 | 0.057 | 1.000 | 1,453.297 |
| | sd(Intercept) | Random | 0.060 | 0.017 | 0.011 | 0.085 | 1.020 | 601.735 |

***Table A9:***

*H6: Previous Turn Acknowledgment Parameter Estimates*

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| STEM | speaker_typetarget | Fixed | 0.115 | 0.074 | -0.027 | 0.265 | 1.000 | 9,000.086 |
| | speaker_typepartner | Fixed | 0.143 | 0.087 | -0.029 | 0.318 | 1.000 | 8,681.533 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.125 | 0.143 | -0.150 | 0.407 | 1.000 | 8,168.481 |
| | sigma | Residual | 0.087 | 0.031 | 0.037 | 0.158 | 1.001 | 4,429.190 |
| | sd(Intercept) | Random | 0.049 | 0.044 | 0.001 | 0.162 | 1.001 | 8,265.730 |
| | sd(Intercept) | Random | 0.048 | 0.044 | 0.001 | 0.162 | 1.000 | 8,374.078 |
| | sd(Intercept) | Random | 0.048 | 0.044 | 0.001 | 0.164 | 1.000 | 8,391.375 |
| Negotiation | speaker_typetarget | Fixed | 0.096 | 0.085 | -0.080 | 0.257 | 1.093 | 48.386 |
| | speaker_typepartner | Fixed | 0.159 | 0.091 | -0.025 | 0.357 | 1.501 | 1,160.402 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.102 | 0.121 | -0.122 | 0.362 | 1.159 | 64.812 |
| | sigma | Residual | 0.046 | 0.053 | 0.002 | 0.184 | 1.553 | 7.038 |
| | sd(Intercept) | Random | 0.079 | 0.058 | 0.004 | 0.229 | 1.504 | 4,476.132 |
| | sd(Intercept) | Random | 0.090 | 0.058 | 0.004 | 0.230 | 1.295 | 127.809 |
| | sd(Intercept) | Random | 0.048 | 0.052 | 0.002 | 0.183 | 1.113 | 24.182 |
| Round Robin (Video Mediated) | speaker_typetarget | Fixed | 0.089 | 0.079 | -0.058 | 0.252 | 1.000 | 5,861.118 |
| | speaker_typepartner | Fixed | 0.067 | 0.079 | -0.082 | 0.232 | 1.000 | 5,935.135 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.108 | 0.111 | -0.115 | 0.321 | 1.001 | 6,826.078 |
| | sigma | Residual | 0.031 | 0.029 | 0.002 | 0.110 | 1.001 | 1,107.157 |
| | sd(Intercept) | Random | 0.051 | 0.049 | 0.002 | 0.184 | 1.000 | 7,020.394 |
| | sd(Intercept) | Random | 0.046 | 0.047 | 0.001 | 0.169 | 1.000 | 6,978.991 |
| | sd(Intercept) | Random | 0.041 | 0.040 | 0.001 | 0.149 | 1.000 | 6,383.639 |
| Round Robin (In Person) | speaker_typetarget | Fixed | 0.141 | 0.036 | 0.071 | 0.211 | 1.000 | 6,118.456 |
| | speaker_typepartner | Fixed | 0.126 | 0.034 | 0.061 | 0.192 | 1.000 | 5,675.167 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.028 | 0.048 | -0.063 | 0.125 | 1.000 | 5,160.977 |
| | sigma | Residual | 0.060 | 0.012 | 0.041 | 0.086 | 1.001 | 2,246.975 |
| | sd(Intercept) | Random | 0.061 | 0.037 | 0.003 | 0.131 | 1.003 | 1,423.641 |
| | sd(Intercept) | Random | 0.064 | 0.037 | 0.003 | 0.130 | 1.001 | 1,202.345 |
| | sd(Intercept) | Random | 0.032 | 0.024 | 0.001 | 0.092 | 1.001 | 2,167.203 |

***Table A10:***

*H6: Previous Turn Agreement Parameter Estimates*

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| New Mothers | speaker_typetarget | Fixed | 0.032 | 0.081 | -0.125 | 0.194 | 1.000 | 7,840.855 |
| | speaker_typepartner | Fixed | 0.220 | 0.081 | 0.064 | 0.386 | 1.001 | 7,667.221 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.105 | 0.095 | -0.086 | 0.288 | 1.000 | 8,765.376 |
| | sigma | Residual | 0.107 | 0.003 | 0.101 | 0.114 | 1.000 | 5,675.510 |
| | sd(Intercept) | Random | 0.019 | 0.010 | 0.001 | 0.038 | 1.002 | 1,735.470 |
| | sd(Intercept) | Random | 0.133 | 0.065 | 0.049 | 0.298 | 1.000 | 7,783.378 |
| | sd(Intercept) | Random | 0.027 | 0.011 | 0.003 | 0.045 | 1.001 | 1,623.009 |
| STEM | speaker_typetarget | Fixed | 0.136 | 0.012 | 0.106 | 0.154 | 1.001 | 6,638.702 |
| | speaker_typepartner | Fixed | 0.129 | 0.009 | 0.112 | 0.146 | 1.000 | 6,911.127 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.040 | 0.012 | 0.017 | 0.064 | 1.000 | 7,246.348 |
| | sigma | Residual | 0.120 | 0.002 | 0.117 | 0.124 | 1.000 | 7,182.942 |
| | sd(Intercept) | Random | 0.011 | 0.007 | 0.001 | 0.026 | 1.002 | 1,709.231 |
| | sd(Intercept) | Random | 0.011 | 0.007 | 0.001 | 0.025 | 1.001 | 1,861.107 |
| | sd(Intercept) | Random | 0.055 | 0.004 | 0.047 | 0.062 | 1.000 | 4,044.216 |
| Middle Eastern Men | speaker_typetarget | Fixed | 0.037 | 0.052 | -0.058 | 0.149 | 1.000 | 6,422.928 |
| | speaker_typepartner | Fixed | 0.020 | 0.075 | -0.121 | 0.178 | 1.000 | 6,755.149 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.207 | 0.057 | 0.086 | 0.311 | 1.000 | 7,584.407 |
| | sigma | Residual | 0.080 | 0.001 | 0.078 | 0.082 | 1.001 | 7,427.647 |
| | sd(Intercept) | Random | 0.004 | 0.003 | 0.000 | 0.010 | 1.000 | 2,851.898 |
| | sd(Intercept) | Random | 0.114 | 0.041 | 0.061 | 0.219 | 1.000 | 6,080.180 |
| | sd(Intercept) | Random | 0.034 | 0.002 | 0.030 | 0.037 | 1.000 | 6,123.005 |
| Negotiation | speaker_typetarget | Fixed | 0.129 | 0.023 | 0.075 | 0.165 | 1.001 | 6,070.438 |
| | speaker_typepartner | Fixed | 0.137 | 0.017 | 0.106 | 0.169 | 1.000 | 7,192.107 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.071 | 0.027 | 0.018 | 0.126 | 1.000 | 6,446.018 |
| | sigma | Residual | 0.123 | 0.004 | 0.115 | 0.132 | 1.000 | 6,228.392 |
| | sd(Intercept) | Random | 0.017 | 0.012 | 0.001 | 0.043 | 1.000 | 3,293.963 |
| | sd(Intercept) | Random | 0.016 | 0.012 | 0.001 | 0.043 | 1.001 | 3,391.839 |
| | sd(Intercept) | Random | 0.048 | 0.011 | 0.026 | 0.068 | 1.001 | 3,417.997 |
| | speaker_typetarget | Fixed | 0.114 | 0.005 | 0.104 | 0.123 | 1.001 | 6,368.671 |

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| Round Robin (Video Mediated) | speaker_typepartner | Fixed | 0.115 | 0.006 | 0.105 | 0.124 | 1.002 | 4,481.689 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.019 | 0.007 | 0.006 | 0.032 | 1.001 | 7,166.215 |
| | sigma | Residual | 0.070 | 0.001 | 0.069 | 0.072 | 1.000 | 6,451.053 |
| | sd(Intercept) | Random | 0.017 | 0.002 | 0.014 | 0.021 | 1.000 | 5,830.842 |
| | sd(Intercept) | Random | 0.005 | 0.003 | 0.000 | 0.010 | 1.001 | 2,284.830 |
| | sd(Intercept) | Random | 0.027 | 0.002 | 0.023 | 0.030 | 1.000 | 4,693.290 |
| Round Robin (In Person) | speaker_typetarget | Fixed | 0.109 | 0.005 | 0.102 | 0.120 | 1.001 | 5,606.063 |
| | speaker_typepartner | Fixed | 0.109 | 0.004 | 0.100 | 0.115 | 1.000 | 6,914.284 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.004 | 0.005 | -0.006 | 0.014 | 1.001 | 7,414.393 |
| | sigma | Residual | 0.070 | 0.001 | 0.069 | 0.072 | 1.000 | 6,867.144 |
| | sd(Intercept) | Random | 0.005 | 0.003 | 0.000 | 0.011 | 1.002 | 1,477.231 |
| | sd(Intercept) | Random | 0.005 | 0.003 | 0.000 | 0.011 | 1.002 | 1,371.501 |
| | sd(Intercept) | Random | 0.031 | 0.002 | 0.028 | 0.034 | 1.000 | 3,938.710 |

***Table A11:***

*H6: Previous Turn Disagreement Parameter Estimates*

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| New Mothers | speaker_typetarget | Fixed | 0.096 | 0.083 | -0.069 | 0.255 | 1.000 | 7,608.447 |
| | speaker_typepartner | Fixed | 0.183 | 0.065 | 0.043 | 0.302 | 1.000 | 6,592.437 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.098 | 0.096 | -0.082 | 0.298 | 1.001 | 8,847.539 |
| | sigma | Residual | 0.105 | 0.003 | 0.099 | 0.111 | 1.001 | 5,649.179 |
| | sd(Intercept) | Random | 0.026 | 0.009 | 0.005 | 0.043 | 1.005 | 1,681.367 |
| | sd(Intercept) | Random | 0.092 | 0.045 | 0.022 | 0.199 | 1.000 | 3,781.837 |
| | sd(Intercept) | Random | 0.064 | 0.007 | 0.051 | 0.077 | 1.001 | 3,222.456 |
| STEM | speaker_typetarget | Fixed | 0.166 | 0.018 | 0.125 | 0.195 | 1.000 | 6,604.268 |
| | speaker_typepartner | Fixed | 0.158 | 0.013 | 0.133 | 0.180 | 1.000 | 7,147.905 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.024 | 0.016 | -0.006 | 0.056 | 1.000 | 7,864.492 |
| | sigma | Residual | 0.123 | 0.002 | 0.119 | 0.127 | 1.000 | 7,301.193 |
| | sd(Intercept) | Random | 0.025 | 0.013 | 0.001 | 0.047 | 1.003 | 617.651 |
| | sd(Intercept) | Random | 0.023 | 0.013 | 0.001 | 0.046 | 1.002 | 718.906 |
| | sd(Intercept) | Random | 0.057 | 0.005 | 0.046 | 0.067 | 1.001 | 4,074.271 |
| Negotiation | speaker_typetarget | Fixed | 0.117 | 0.023 | 0.073 | 0.162 | 1.000 | 7,137.393 |
| | speaker_typepartner | Fixed | 0.130 | 0.022 | 0.086 | 0.172 | 1.000 | 7,327.405 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.079 | 0.036 | 0.010 | 0.152 | 1.000 | 7,908.769 |
| | sigma | Residual | 0.119 | 0.006 | 0.108 | 0.131 | 1.000 | 7,654.249 |
| | sd(Intercept) | Random | 0.028 | 0.018 | 0.001 | 0.066 | 1.001 | 2,643.628 |
| | sd(Intercept) | Random | 0.028 | 0.018 | 0.001 | 0.066 | 1.002 | 2,910.780 |
| | sd(Intercept) | Random | 0.017 | 0.013 | 0.001 | 0.047 | 1.001 | 4,590.379 |
| Round Robin (Video Mediated) | speaker_typetarget | Fixed | 0.112 | 0.007 | 0.098 | 0.123 | 1.002 | 2,920.312 |
| | speaker_typepartner | Fixed | 0.115 | 0.010 | 0.103 | 0.132 | 1.003 | 1,261.979 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.006 | 0.009 | -0.012 | 0.024 | 1.001 | 2,849.805 |
| | sigma | Residual | 0.071 | 0.001 | 0.069 | 0.074 | 1.002 | 4,890.083 |
| | sd(Intercept) | Random | 0.005 | 0.003 | 0.000 | 0.012 | 1.001 | 2,155.477 |
| | sd(Intercept) | Random | 0.004 | 0.003 | 0.000 | 0.010 | 1.003 | 2,541.505 |
| | sd(Intercept) | Random | 0.033 | 0.002 | 0.029 | 0.038 | 1.001 | 4,684.505 |
| | speaker_typetarget | Fixed | 0.113 | 0.005 | 0.104 | 0.121 | 1.001 | 5,728.398 |

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| Round Robin (In Person) | speaker_typepartner | Fixed | 0.112 | 0.006 | 0.101 | 0.122 | 1.001 | 6,499.440 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.000 | 0.007 | -0.013 | 0.014 | 1.000 | 8,071.651 |
| | sigma | Residual | 0.072 | 0.001 | 0.070 | 0.073 | 1.000 | 6,015.516 |
| | sd(Intercept) | Random | 0.007 | 0.004 | 0.000 | 0.015 | 1.003 | 1,137.209 |
| | sd(Intercept) | Random | 0.007 | 0.004 | 0.000 | 0.016 | 1.001 | 1,038.719 |
| | sd(Intercept) | Random | 0.033 | 0.002 | 0.028 | 0.037 | 1.001 | 2,638.038 |

***Table A12:***

*H6: Previous Turn Functional Parameter Estimates*

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| STEM | speaker_typetarget | Fixed | 0.160 | 0.019 | 0.126 | 0.198 | 1.000 | 7,836.751 |
| | speaker_typepartner | Fixed | 0.146 | 0.017 | 0.113 | 0.179 | 1.000 | 8,293.805 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.036 | 0.023 | -0.011 | 0.081 | 1.000 | 8,039.952 |
| | sigma | Residual | 0.129 | 0.003 | 0.124 | 0.136 | 1.001 | 6,252.878 |
| | sd(Intercept) | Random | 0.023 | 0.014 | 0.001 | 0.049 | 1.003 | 1,901.606 |
| | sd(Intercept) | Random | 0.023 | 0.014 | 0.001 | 0.049 | 1.002 | 1,810.452 |
| | sd(Intercept) | Random | 0.052 | 0.010 | 0.031 | 0.070 | 1.000 | 2,859.896 |
| Negotiation | speaker_typetarget | Fixed | 0.111 | 0.034 | 0.045 | 0.175 | 1.000 | 8,812.158 |
| | speaker_typepartner | Fixed | 0.112 | 0.032 | 0.049 | 0.174 | 1.000 | 8,058.594 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.120 | 0.052 | 0.020 | 0.223 | 1.000 | 8,701.521 |
| | sigma | Residual | 0.116 | 0.010 | 0.098 | 0.136 | 1.000 | 7,448.373 |
| | sd(Intercept) | Random | 0.029 | 0.021 | 0.001 | 0.077 | 1.000 | 6,211.094 |
| | sd(Intercept) | Random | 0.029 | 0.020 | 0.001 | 0.077 | 1.000 | 6,551.121 |
| | sd(Intercept) | Random | 0.026 | 0.019 | 0.001 | 0.068 | 1.000 | 5,937.990 |
| Round Robin (Video Mediated) | speaker_typetarget | Fixed | 0.118 | 0.011 | 0.095 | 0.137 | 1.000 | 6,950.182 |
| | speaker_typepartner | Fixed | 0.130 | 0.010 | 0.109 | 0.148 | 1.000 | 7,258.911 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | -0.002 | 0.015 | -0.031 | 0.028 | 1.000 | 8,426.471 |
| | sigma | Residual | 0.070 | 0.003 | 0.065 | 0.075 | 1.001 | 4,512.207 |
| | sd(Intercept) | Random | 0.028 | 0.008 | 0.009 | 0.041 | 1.001 | 1,916.013 |
| | sd(Intercept) | Random | 0.009 | 0.006 | 0.000 | 0.024 | 1.002 | 3,077.935 |
| | sd(Intercept) | Random | 0.027 | 0.009 | 0.006 | 0.041 | 1.001 | 1,746.587 |
| Round Robin (In Person) | speaker_typetarget | Fixed | 0.098 | 0.007 | 0.086 | 0.110 | 1.000 | 7,534.008 |
| | speaker_typepartner | Fixed | 0.096 | 0.007 | 0.083 | 0.109 | 1.001 | 7,266.095 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.019 | 0.010 | 0.001 | 0.039 | 1.001 | 9,034.956 |
| | sigma | Residual | 0.069 | 0.002 | 0.066 | 0.072 | 1.000 | 6,521.621 |
| | sd(Intercept) | Random | 0.010 | 0.006 | 0.000 | 0.023 | 1.002 | 1,513.190 |
| | sd(Intercept) | Random | 0.010 | 0.006 | 0.000 | 0.023 | 1.002 | 1,414.789 |
| | sd(Intercept) | Random | 0.029 | 0.004 | 0.020 | 0.037 | 1.002 | 2,156.902 |

***Table A13:***

*H6: Previous Turn Inform Parameter Estimates*

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| New Mothers | speaker_typetarget | Fixed | 0.106 | 0.097 | -0.086 | 0.296 | 1.001 | 5,777.308 |
| | speaker_typepartner | Fixed | 0.129 | 0.047 | 0.055 | 0.241 | 1.001 | 3,080.624 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.179 | 0.087 | -0.030 | 0.303 | 1.000 | 3,155.181 |
| | sigma | Residual | 0.104 | 0.001 | 0.102 | 0.106 | 1.000 | 7,746.482 |
| | sd(Intercept) | Random | 0.027 | 0.003 | 0.021 | 0.034 | 1.001 | 4,246.278 |
| | sd(Intercept) | Random | 0.042 | 0.017 | 0.021 | 0.085 | 1.001 | 4,282.368 |
| | sd(Intercept) | Random | 0.036 | 0.002 | 0.031 | 0.041 | 1.000 | 4,412.568 |
| STEM | speaker_typetarget | Fixed | 0.151 | 0.021 | 0.094 | 0.177 | 1.000 | 6,845.625 |
| | speaker_typepartner | Fixed | 0.156 | 0.010 | 0.136 | 0.173 | 1.000 | 7,014.406 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.029 | 0.011 | 0.007 | 0.053 | 1.000 | 7,488.135 |
| | sigma | Residual | 0.128 | 0.001 | 0.126 | 0.130 | 1.000 | 6,651.455 |
| | sd(Intercept) | Random | 0.017 | 0.009 | 0.001 | 0.032 | 1.007 | 609.811 |
| | sd(Intercept) | Random | 0.017 | 0.009 | 0.001 | 0.031 | 1.004 | 657.962 |
| | sd(Intercept) | Random | 0.053 | 0.002 | 0.048 | 0.058 | 1.001 | 5,482.496 |
| Negotiation | speaker_typetarget | Fixed | 0.144 | 0.028 | 0.101 | 0.213 | 1.000 | 6,998.228 |
| | speaker_typepartner | Fixed | 0.121 | 0.013 | 0.093 | 0.145 | 1.000 | 7,665.107 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.107 | 0.022 | 0.066 | 0.153 | 1.000 | 7,181.631 |
| | sigma | Residual | 0.133 | 0.003 | 0.128 | 0.139 | 1.001 | 6,139.299 |
| | sd(Intercept) | Random | 0.014 | 0.009 | 0.001 | 0.033 | 1.000 | 3,697.511 |
| | sd(Intercept) | Random | 0.014 | 0.009 | 0.001 | 0.033 | 1.001 | 3,513.594 |
| | sd(Intercept) | Random | 0.039 | 0.008 | 0.024 | 0.054 | 1.001 | 4,058.169 |
| Round Robin (Video Mediated) | speaker_typetarget | Fixed | 0.117 | 0.004 | 0.110 | 0.123 | 1.000 | 7,289.398 |
| | speaker_typepartner | Fixed | 0.119 | 0.003 | 0.113 | 0.125 | 1.000 | 7,589.124 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | -0.001 | 0.004 | -0.008 | 0.007 | 1.000 | 7,596.995 |
| | sigma | Residual | 0.070 | 0.001 | 0.069 | 0.071 | 1.001 | 6,570.917 |
| | sd(Intercept) | Random | 0.015 | 0.002 | 0.011 | 0.018 | 1.001 | 4,654.507 |
| | sd(Intercept) | Random | 0.005 | 0.003 | 0.000 | 0.010 | 1.001 | 1,497.680 |
| | sd(Intercept) | Random | 0.031 | 0.001 | 0.029 | 0.034 | 1.001 | 5,235.918 |
| | speaker_typetarget | Fixed | 0.110 | 0.004 | 0.105 | 0.117 | 1.001 | 6,885.250 |

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| Round Robin (In Person) | speaker_typepartner | Fixed | 0.109 | 0.004 | 0.099 | 0.115 | 1.000 | 7,211.584 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.005 | 0.004 | -0.002 | 0.012 | 1.000 | 8,157.864 |
| | sigma | Residual | 0.072 | 0.000 | 0.072 | 0.073 | 1.000 | 7,704.784 |
| | sd(Intercept) | Random | 0.006 | 0.003 | 0.000 | 0.011 | 1.014 | 440.797 |
| | sd(Intercept) | Random | 0.006 | 0.003 | 0.000 | 0.011 | 1.028 | 285.034 |
| | sd(Intercept) | Random | 0.033 | 0.001 | 0.031 | 0.035 | 1.001 | 4,075.162 |

***Table A14:***

*H6: Previous Turn Question Parameter Estimates*

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| New Mothers | speaker_typetarget | Fixed | 0.159 | 0.072 | 0.015 | 0.301 | 1.000 | 8,285.820 |
| | speaker_typepartner | Fixed | 0.153 | 0.064 | 0.025 | 0.281 | 1.000 | 8,306.701 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.072 | 0.097 | -0.115 | 0.262 | 1.000 | 8,973.813 |
| | sigma | Residual | 0.092 | 0.002 | 0.088 | 0.096 | 1.000 | 6,484.341 |
| | sd(Intercept) | Random | 0.026 | 0.004 | 0.018 | 0.034 | 1.000 | 5,447.588 |
| | sd(Intercept) | Random | 0.083 | 0.030 | 0.043 | 0.159 | 1.000 | 7,509.968 |
| | sd(Intercept) | Random | 0.045 | 0.004 | 0.037 | 0.052 | 1.000 | 4,758.088 |
| STEM | speaker_typetarget | Fixed | 0.176 | 0.013 | 0.150 | 0.200 | 1.000 | 7,022.516 |
| | speaker_typepartner | Fixed | 0.187 | 0.014 | 0.161 | 0.213 | 1.000 | 6,900.003 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.011 | 0.016 | -0.018 | 0.045 | 1.000 | 7,953.669 |
| | sigma | Residual | 0.139 | 0.002 | 0.135 | 0.143 | 1.000 | 6,884.291 |
| | sd(Intercept) | Random | 0.015 | 0.009 | 0.001 | 0.031 | 1.006 | 1,041.329 |
| | sd(Intercept) | Random | 0.015 | 0.009 | 0.001 | 0.031 | 1.002 | 1,375.502 |
| | sd(Intercept) | Random | 0.053 | 0.005 | 0.043 | 0.063 | 1.000 | 4,302.676 |
| Negotiation | speaker_typetarget | Fixed | 0.128 | 0.022 | 0.083 | 0.170 | 1.000 | 6,985.892 |
| | speaker_typepartner | Fixed | 0.128 | 0.019 | 0.091 | 0.165 | 1.000 | 6,848.218 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.086 | 0.032 | 0.023 | 0.149 | 1.000 | 6,694.873 |
| | sigma | Residual | 0.132 | 0.005 | 0.122 | 0.143 | 1.000 | 5,720.524 |
| | sd(Intercept) | Random | 0.018 | 0.012 | 0.001 | 0.044 | 1.001 | 3,819.059 |
| | sd(Intercept) | Random | 0.018 | 0.012 | 0.001 | 0.045 | 1.000 | 3,563.555 |
| | sd(Intercept) | Random | 0.036 | 0.014 | 0.004 | 0.061 | 1.002 | 2,549.326 |
| Round Robin (Video Mediated) | speaker_typetarget | Fixed | 0.111 | 0.004 | 0.103 | 0.118 | 1.000 | 7,884.948 |
| | speaker_typepartner | Fixed | 0.113 | 0.004 | 0.105 | 0.120 | 1.000 | 7,172.410 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.015 | 0.005 | 0.005 | 0.026 | 1.000 | 8,510.533 |
| | sigma | Residual | 0.078 | 0.001 | 0.076 | 0.080 | 1.000 | 6,636.286 |
| | sd(Intercept) | Random | 0.013 | 0.002 | 0.008 | 0.018 | 1.000 | 3,521.876 |
| | sd(Intercept) | Random | 0.005 | 0.003 | 0.000 | 0.010 | 1.002 | 2,333.742 |
| | sd(Intercept) | Random | 0.033 | 0.002 | 0.029 | 0.036 | 1.001 | 4,332.616 |
| | speaker_typetarget | Fixed | 0.109 | 0.004 | 0.103 | 0.117 | 1.000 | 5,697.056 |

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| Round Robin (In Person) | speaker_typepartner | Fixed | 0.107 | 0.005 | 0.096 | 0.115 | 1.001 | 6,395.129 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.001 | 0.005 | -0.010 | 0.011 | 1.000 | 6,383.795 |
| | sigma | Residual | 0.074 | 0.001 | 0.072 | 0.075 | 1.000 | 6,882.530 |
| | sd(Intercept) | Random | 0.006 | 0.004 | 0.000 | 0.013 | 1.004 | 704.073 |
| | sd(Intercept) | Random | 0.007 | 0.004 | 0.000 | 0.013 | 1.008 | 675.049 |
| | sd(Intercept) | Random | 0.032 | 0.001 | 0.029 | 0.035 | 1.000 | 3,220.774 |

***Table A15:***

*H6: Previous Turn Suggestion Parameter Estimates*

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| STEM | speaker_typetarget | Fixed | 0.175 | 0.030 | 0.116 | 0.232 | 1.000 | 8,825.489 |
| | speaker_typepartner | Fixed | 0.147 | 0.033 | 0.081 | 0.211 | 1.000 | 8,460.692 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.023 | 0.039 | -0.053 | 0.100 | 1.000 | 8,408.998 |
| | sigma | Residual | 0.123 | 0.006 | 0.111 | 0.136 | 1.000 | 7,127.491 |
| | sd(Intercept) | Random | 0.034 | 0.022 | 0.001 | 0.080 | 1.000 | 3,420.388 |
| | sd(Intercept) | Random | 0.035 | 0.023 | 0.002 | 0.082 | 1.000 | 3,160.659 |
| | sd(Intercept) | Random | 0.046 | 0.024 | 0.003 | 0.089 | 1.001 | 2,984.393 |
| Negotiation | speaker_typetarget | Fixed | 0.106 | 0.064 | -0.024 | 0.229 | 1.000 | 8,349.088 |
| | speaker_typepartner | Fixed | 0.138 | 0.063 | 0.014 | 0.262 | 1.000 | 8,341.227 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.097 | 0.095 | -0.087 | 0.281 | 1.000 | 8,426.916 |
| | sigma | Residual | 0.108 | 0.021 | 0.075 | 0.156 | 1.000 | 6,869.886 |
| | sd(Intercept) | Random | 0.055 | 0.043 | 0.002 | 0.160 | 1.000 | 7,080.420 |
| | sd(Intercept) | Random | 0.054 | 0.043 | 0.002 | 0.156 | 1.000 | 6,773.084 |
| | sd(Intercept) | Random | 0.051 | 0.040 | 0.002 | 0.149 | 1.000 | 7,302.621 |
| Round Robin (Video Mediated) | speaker_typetarget | Fixed | 0.102 | 0.025 | 0.053 | 0.149 | 1.000 | 7,954.350 |
| | speaker_typepartner | Fixed | 0.113 | 0.019 | 0.075 | 0.150 | 1.000 | 8,692.341 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.043 | 0.033 | -0.024 | 0.109 | 1.000 | 8,415.470 |
| | sigma | Residual | 0.071 | 0.006 | 0.060 | 0.083 | 1.000 | 5,788.944 |
| | sd(Intercept) | Random | 0.016 | 0.011 | 0.001 | 0.041 | 1.000 | 4,982.400 |
| | sd(Intercept) | Random | 0.017 | 0.011 | 0.001 | 0.040 | 1.000 | 4,857.979 |
| | sd(Intercept) | Random | 0.019 | 0.013 | 0.001 | 0.047 | 1.001 | 3,933.147 |
| Round Robin (In Person) | speaker_typetarget | Fixed | 0.126 | 0.017 | 0.094 | 0.160 | 1.000 | 7,595.208 |
| | speaker_typepartner | Fixed | 0.132 | 0.017 | 0.099 | 0.164 | 1.001 | 7,983.805 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | -0.026 | 0.027 | -0.080 | 0.026 | 1.001 | 8,156.159 |
| | sigma | Residual | 0.067 | 0.005 | 0.059 | 0.077 | 1.001 | 5,492.532 |
| | sd(Intercept) | Random | 0.025 | 0.016 | 0.001 | 0.055 | 1.005 | 1,329.730 |
| | sd(Intercept) | Random | 0.025 | 0.016 | 0.001 | 0.056 | 1.005 | 1,299.679 |
| | sd(Intercept) | Random | 0.025 | 0.015 | 0.001 | 0.054 | 1.002 | 1,294.676 |

# APPENDIX B

## STUDY 3 DIALOGUE ACT MODEL FIT STATISTICS

Beginning with Table A2, each table shows the estimates for all parameters within a single dialogue act. These retain the variable names from the original brms code, which are explained here and at the beginning of Appendix B.

**speaker_typetarget** and **speaker_typepartner**: Mean cosine similarity for target or interactant speech. Because the model drops the intercept (the `0 +` syntax), this is the mean, not a difference from some reference group.

**I(percent_of_chapter_at_utterance_start)^2**: This is a labeling artifact from the HSGP approximation. The actual variable underneath is the normalized position of each utterance in its chapter (0 = start, 1 = end) — i.e., the time axis that the Gaussian process is smoothing over.

**sigma**: Residual standard deviation; whatever observation-level noise is left after the fixed effects, GP, and random effects.

**sd(Intercept)** — shows up three times per study because there are three nested grouping factors. *Perceiver* (or target, in Appendix B): How much individual perceivers differ from each other in baseline cosine similarity. *Target*: Between-target variability in the sample. *Chapter*: How much cosine similarity bounces around from chapter to chapter within the same target.

***Table B1:***

*Convergence Diagnostics for Study 3 Target Dialogue Act Models*

| Hypothesis | Dialogue Act | Study | Max Rhat | Min ESS | Divergent |
|---|---|---|---|---|---|
| H5 | Acknowledgment | Middle Eastern Men | 3.598 | 4 | 0 |
| | | Negotiation | 1.001 | 6,668 | 118 |
| | | Round Robin (Video Mediated) | 1.161 | 18 | 743 |
| | Agreement | Middle Eastern Men | 1.058 | 100 | 0 |
| | | Negotiation | 1.001 | 5,329 | 1 |
| | | Round Robin (Video Mediated) | 1.003 | 1,438 | 105 |
| | Disagreement | Negotiation | 1.002 | 2,309 | 0 |
| | | Round Robin (Video Mediated) | 1.005 | 1,558 | 37 |
| | Functional | Negotiation | 1.001 | 1,995 | 0 |
| | | Round Robin (Video Mediated) | 1.001 | 2,031 | 0 |
| | Inform | Negotiation | 1.003 | 2,215 | 4 |
| | | Round Robin (Video Mediated) | 1.007 | 976 | 38 |
| | Question | Middle Eastern Men | 1.232 | 14 | 0 |
| | | Negotiation | 1.002 | 2,087 | 0 |
| | | Round Robin (Video Mediated) | 1.001 | 2,532 | 14 |
| | Suggestion | Negotiation | 1.001 | 5,580 | 40 |
| | | Round Robin (Video Mediated) | 1.002 | 1,548 | 0 |
| H6 | Acknowledgment | Negotiation | 1.003 | 2,001 | 334 |
| | | Round Robin (Video Mediated) | 1.082 | 34 | 278 |
| | Agreement | Middle Eastern Men | 1.013 | 296 | 0 |
| | | Negotiation | 1.000 | 5,053 | 4 |
| | | Round Robin (Video Mediated) | 1.003 | 1,676 | 44 |
| | Disagreement | Middle Eastern Men | 1.001 | 6,900 | 0 |

| Hypothesis | Dialogue Act | Study | Max Rhat | Min ESS | Divergent |
|---|---|---|---|---|---|
| | | Negotiation | 1.001 | 3,135 | 0 |
| | | Round Robin (Video Mediated) | 1.001 | 1,815 | 18 |
| | Functional | Middle Eastern Men | 1.058 | 74 | 0 |
| | | Negotiation | 1.000 | 5,464 | 0 |
| | | Round Robin (Video Mediated) | 1.002 | 1,833 | 0 |
| | Inform | Negotiation | 1.001 | 2,128 | 1 |
| | Question | Negotiation | 1.002 | 2,475 | 0 |
| | | Round Robin (Video Mediated) | 1.005 | 1,444 | 2 |
| | Suggestion | Negotiation | 1.000 | 7,792 | 0 |
| | | Round Robin (Video Mediated) | 1.002 | 3,426 | 0 |

***Table B2:***

*RQ1: Acknowledgment Parameter Estimates*

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| Middle Eastern Men | speaker_typetarget | Fixed | 0.180 | 0.048 | 0.125 | 0.248 | 3.323 | 4.434 |
| | speaker_typepartner | Fixed | 0.196 | 0.069 | 0.124 | 0.317 | 3.406 | 4.403 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | -0.154 | 0.173 | -0.448 | 0.027 | 3.598 | 4.363 |
| | sigma | Residual | 0.000 | 0.000 | 0.000 | 0.000 | 1.124 | 25.818 |
| | sd(Intercept) | Random | 0.000 | 0.000 | 0.000 | 0.000 | 1.601 | 6.925 |
| | sd(Intercept) | Random | 0.049 | 0.045 | 0.019 | 0.129 | 3.110 | 4.559 |
| | sd(Intercept) | Random | 0.000 | 0.000 | 0.000 | 0.000 | 1.087 | 42.272 |
| Negotiation | speaker_typetarget | Fixed | 0.097 | 0.059 | -0.021 | 0.215 | 1.000 | 8,001.562 |
| | speaker_typepartner | Fixed | 0.060 | 0.052 | -0.042 | 0.167 | 1.000 | 6,667.912 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.045 | 0.074 | -0.097 | 0.207 | 1.000 | 7,215.981 |
| | sigma | Residual | 0.026 | 0.021 | 0.002 | 0.077 | 1.001 | 1,205.876 |
| | sd(Intercept) | Random | 0.033 | 0.028 | 0.001 | 0.103 | 1.001 | 6,842.239 |
| | sd(Intercept) | Random | 0.033 | 0.028 | 0.001 | 0.105 | 1.000 | 7,309.735 |
| | sd(Intercept) | Random | 0.028 | 0.024 | 0.001 | 0.088 | 1.000 | 6,925.119 |
| Round Robin (Video Mediated) | speaker_typetarget | Fixed | 0.132 | 0.033 | 0.072 | 0.208 | 1.029 | 170.776 |
| | speaker_typepartner | Fixed | 0.106 | 0.025 | 0.059 | 0.154 | 1.048 | 59.283 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.002 | 0.035 | -0.076 | 0.070 | 1.076 | 252.221 |
| | sigma | Residual | 0.010 | 0.012 | 0.000 | 0.045 | 1.111 | 24.929 |
| | sd(Intercept) | Random | 0.035 | 0.019 | 0.002 | 0.071 | 1.103 | 29.595 |
| | sd(Intercept) | Random | 0.056 | 0.026 | 0.004 | 0.105 | 1.123 | 24.475 |
| | sd(Intercept) | Random | 0.040 | 0.027 | 0.003 | 0.097 | 1.161 | 18.490 |

***Table B3:***

*RQ1: Agreement Parameter Estimates*

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| Middle Eastern Men | speaker_typetarget | Fixed | 0.129 | 0.139 | -0.155 | 0.399 | 1.038 | 247.242 |
| | speaker_typepartner | Fixed | 0.107 | 0.048 | 0.013 | 0.204 | 1.040 | 99.631 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.063 | 0.087 | -0.096 | 0.252 | 1.030 | 109.137 |
| | sigma | Residual | 0.063 | 0.001 | 0.062 | 0.064 | 1.015 | 266.842 |
| | sd(Intercept) | Random | 0.001 | 0.000 | 0.000 | 0.002 | 1.011 | 211.423 |
| | sd(Intercept) | Random | 0.044 | 0.016 | 0.023 | 0.086 | 1.010 | 204.530 |
| | sd(Intercept) | Random | 0.001 | 0.001 | 0.000 | 0.003 | 1.058 | 106.075 |
| Negotiation | speaker_typetarget | Fixed | 0.113 | 0.010 | 0.093 | 0.133 | 1.000 | 7,295.588 |
| | speaker_typepartner | Fixed | 0.109 | 0.010 | 0.090 | 0.128 | 1.000 | 6,826.556 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.055 | 0.014 | 0.029 | 0.082 | 1.000 | 8,162.473 |
| | sigma | Residual | 0.103 | 0.003 | 0.098 | 0.109 | 1.000 | 6,692.233 |
| | sd(Intercept) | Random | 0.006 | 0.005 | 0.000 | 0.017 | 1.001 | 5,985.091 |
| | sd(Intercept) | Random | 0.006 | 0.005 | 0.000 | 0.017 | 1.000 | 5,524.529 |
| | sd(Intercept) | Random | 0.044 | 0.006 | 0.033 | 0.055 | 1.000 | 5,329.150 |
| Round Robin (Video Mediated) | speaker_typetarget | Fixed | 0.141 | 0.005 | 0.134 | 0.148 | 1.001 | 5,088.198 |
| | speaker_typepartner | Fixed | 0.135 | 0.011 | 0.102 | 0.149 | 1.002 | 2,818.577 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.004 | 0.004 | -0.005 | 0.013 | 1.001 | 6,215.385 |
| | sigma | Residual | 0.058 | 0.000 | 0.057 | 0.059 | 1.000 | 7,332.631 |
| | sd(Intercept) | Random | 0.003 | 0.002 | 0.000 | 0.008 | 1.003 | 1,438.000 |
| | sd(Intercept) | Random | 0.017 | 0.001 | 0.014 | 0.020 | 1.001 | 4,062.858 |
| | sd(Intercept) | Random | 0.036 | 0.001 | 0.034 | 0.038 | 1.001 | 3,930.272 |

***Table B4:***

*RQ1: Disagreement Parameter Estimates*

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| Negotiation | speaker_typetarget | Fixed | 0.114 | 0.027 | 0.063 | 0.170 | 1.000 | 8,505.444 |
| | speaker_typepartner | Fixed | 0.125 | 0.030 | 0.070 | 0.185 | 1.000 | 8,371.134 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.129 | 0.035 | 0.061 | 0.200 | 1.000 | 8,753.877 |
| | sigma | Residual | 0.134 | 0.009 | 0.117 | 0.152 | 1.001 | 3,395.301 |
| | sd(Intercept) | Random | 0.038 | 0.025 | 0.002 | 0.094 | 1.000 | 3,283.501 |
| | sd(Intercept) | Random | 0.038 | 0.026 | 0.002 | 0.094 | 1.001 | 3,326.160 |
| | sd(Intercept) | Random | 0.061 | 0.023 | 0.010 | 0.102 | 1.002 | 2,309.175 |
| Round Robin (Video Mediated) | speaker_typetarget | Fixed | 0.118 | 0.005 | 0.110 | 0.128 | 1.001 | 7,286.006 |
| | speaker_typepartner | Fixed | 0.120 | 0.005 | 0.110 | 0.130 | 1.001 | 5,775.578 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.008 | 0.006 | -0.003 | 0.021 | 1.000 | 6,549.857 |
| | sigma | Residual | 0.078 | 0.001 | 0.076 | 0.080 | 1.001 | 5,873.151 |
| | sd(Intercept) | Random | 0.008 | 0.003 | 0.001 | 0.014 | 1.005 | 1,558.030 |
| | sd(Intercept) | Random | 0.010 | 0.003 | 0.002 | 0.016 | 1.004 | 1,692.295 |
| | sd(Intercept) | Random | 0.037 | 0.002 | 0.032 | 0.041 | 1.000 | 4,112.441 |

***Table B5:***

*RQ1: Functional Parameter Estimates*

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| Negotiation | speaker_typetarget | Fixed | 0.167 | 0.029 | 0.111 | 0.226 | 1.000 | 7,772.324 |
| | speaker_typepartner | Fixed | 0.144 | 0.030 | 0.084 | 0.203 | 1.001 | 8,142.817 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.082 | 0.042 | 0.001 | 0.164 | 1.000 | 7,903.008 |
| | sigma | Residual | 0.133 | 0.013 | 0.107 | 0.158 | 1.001 | 2,638.119 |
| | sd(Intercept) | Random | 0.024 | 0.018 | 0.001 | 0.067 | 1.000 | 4,994.770 |
| | sd(Intercept) | Random | 0.023 | 0.018 | 0.001 | 0.066 | 1.001 | 5,612.955 |
| | sd(Intercept) | Random | 0.058 | 0.031 | 0.003 | 0.114 | 1.001 | 1,995.423 |
| Round Robin (Video Mediated) | speaker_typetarget | Fixed | 0.139 | 0.008 | 0.123 | 0.154 | 1.000 | 6,874.952 |
| | speaker_typepartner | Fixed | 0.142 | 0.008 | 0.127 | 0.156 | 1.000 | 6,842.642 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | -0.001 | 0.011 | -0.023 | 0.021 | 1.000 | 8,122.695 |
| | sigma | Residual | 0.083 | 0.003 | 0.078 | 0.088 | 1.000 | 4,772.635 |
| | sd(Intercept) | Random | 0.012 | 0.007 | 0.001 | 0.025 | 1.001 | 2,030.531 |
| | sd(Intercept) | Random | 0.009 | 0.006 | 0.000 | 0.022 | 1.001 | 2,312.165 |
| | sd(Intercept) | Random | 0.050 | 0.005 | 0.041 | 0.059 | 1.000 | 3,338.348 |

***Table B6:***

*RQ1: Inform Parameter Estimates*

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| Negotiation | speaker_typetarget | Fixed | 0.151 | 0.017 | 0.119 | 0.182 | 1.000 | 6,523.206 |
| | speaker_typepartner | Fixed | 0.146 | 0.013 | 0.122 | 0.170 | 1.000 | 6,261.832 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.100 | 0.015 | 0.070 | 0.131 | 1.000 | 7,693.862 |
| | sigma | Residual | 0.147 | 0.003 | 0.141 | 0.153 | 1.001 | 7,100.169 |
| | sd(Intercept) | Random | 0.019 | 0.011 | 0.001 | 0.042 | 1.003 | 2,353.060 |
| | sd(Intercept) | Random | 0.019 | 0.011 | 0.001 | 0.042 | 1.003 | 2,214.622 |
| | sd(Intercept) | Random | 0.052 | 0.007 | 0.038 | 0.067 | 1.000 | 5,015.922 |
| Round Robin (Video Mediated) | speaker_typetarget | Fixed | 0.135 | 0.003 | 0.131 | 0.141 | 1.000 | 6,029.337 |
| | speaker_typepartner | Fixed | 0.133 | 0.004 | 0.123 | 0.139 | 1.002 | 4,492.915 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.000 | 0.003 | -0.005 | 0.006 | 1.001 | 6,985.470 |
| | sigma | Residual | 0.077 | 0.000 | 0.076 | 0.078 | 1.000 | 7,514.562 |
| | sd(Intercept) | Random | 0.007 | 0.002 | 0.002 | 0.011 | 1.007 | 975.782 |
| | sd(Intercept) | Random | 0.009 | 0.002 | 0.005 | 0.012 | 1.004 | 1,738.033 |
| | sd(Intercept) | Random | 0.038 | 0.001 | 0.036 | 0.040 | 1.001 | 5,023.575 |

***Table B7:***

*RQ1: Question Parameter Estimates*

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| Middle Eastern Men | speaker_typetarget | Fixed | 0.204 | 0.123 | -0.013 | 0.481 | 1.232 | 13.760 |
| | speaker_typepartner | Fixed | 0.117 | 0.150 | -0.158 | 0.433 | 1.021 | 108.310 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.099 | 0.139 | -0.181 | 0.374 | 1.054 | 77.081 |
| | sigma | Residual | 0.084 | 0.001 | 0.083 | 0.086 | 1.057 | 73.198 |
| | sd(Intercept) | Random | 0.001 | 0.001 | 0.000 | 0.002 | 1.099 | 29.235 |
| | sd(Intercept) | Random | 0.051 | 0.011 | 0.033 | 0.080 | 1.149 | 19.673 |
| | sd(Intercept) | Random | 0.001 | 0.001 | 0.000 | 0.003 | 1.193 | 15.241 |
| Negotiation | speaker_typetarget | Fixed | 0.142 | 0.019 | 0.105 | 0.180 | 1.000 | 6,930.446 |
| | speaker_typepartner | Fixed | 0.133 | 0.020 | 0.094 | 0.171 | 1.000 | 6,918.001 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.079 | 0.027 | 0.026 | 0.130 | 1.001 | 8,346.081 |
| | sigma | Residual | 0.131 | 0.006 | 0.119 | 0.145 | 1.001 | 2,924.661 |
| | sd(Intercept) | Random | 0.013 | 0.010 | 0.001 | 0.036 | 1.002 | 4,834.590 |
| | sd(Intercept) | Random | 0.013 | 0.010 | 0.000 | 0.036 | 1.000 | 5,257.040 |
| | sd(Intercept) | Random | 0.059 | 0.016 | 0.022 | 0.087 | 1.000 | 2,087.161 |
| Round Robin (Video Mediated) | speaker_typetarget | Fixed | 0.135 | 0.008 | 0.121 | 0.149 | 1.000 | 5,723.998 |
| | speaker_typepartner | Fixed | 0.128 | 0.011 | 0.106 | 0.148 | 1.000 | 5,500.767 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.016 | 0.010 | -0.001 | 0.038 | 1.001 | 6,065.406 |
| | sigma | Residual | 0.089 | 0.001 | 0.087 | 0.092 | 1.001 | 5,945.414 |
| | sd(Intercept) | Random | 0.004 | 0.003 | 0.000 | 0.011 | 1.001 | 2,532.091 |
| | sd(Intercept) | Random | 0.014 | 0.003 | 0.007 | 0.019 | 1.001 | 2,872.975 |
| | sd(Intercept) | Random | 0.038 | 0.002 | 0.033 | 0.042 | 1.001 | 4,259.952 |

***Table B8:***

*RQ1: Suggestion Parameter Estimates*

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| Negotiation | speaker_typetarget | Fixed | 0.154 | 0.074 | 0.007 | 0.298 | 1.000 | 7,710.300 |
| | speaker_typepartner | Fixed | 0.161 | 0.076 | 0.006 | 0.308 | 1.000 | 9,306.978 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.080 | 0.093 | -0.097 | 0.265 | 1.000 | 9,165.209 |
| | sigma | Residual | 0.130 | 0.041 | 0.045 | 0.221 | 1.001 | 2,268.177 |
| | sd(Intercept) | Random | 0.047 | 0.039 | 0.002 | 0.146 | 1.000 | 7,632.362 |
| | sd(Intercept) | Random | 0.048 | 0.041 | 0.002 | 0.154 | 1.000 | 7,797.500 |
| | sd(Intercept) | Random | 0.041 | 0.035 | 0.001 | 0.129 | 1.001 | 5,579.856 |
| Round Robin (Video Mediated) | speaker_typetarget | Fixed | 0.158 | 0.019 | 0.121 | 0.194 | 1.000 | 7,765.298 |
| | speaker_typepartner | Fixed | 0.147 | 0.016 | 0.112 | 0.178 | 1.000 | 6,565.244 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | -0.029 | 0.026 | -0.077 | 0.024 | 1.000 | 7,083.599 |
| | sigma | Residual | 0.084 | 0.006 | 0.071 | 0.095 | 1.001 | 2,316.279 |
| | sd(Intercept) | Random | 0.012 | 0.009 | 0.000 | 0.033 | 1.001 | 4,464.299 |
| | sd(Intercept) | Random | 0.025 | 0.012 | 0.002 | 0.047 | 1.001 | 2,530.208 |
| | sd(Intercept) | Random | 0.024 | 0.015 | 0.001 | 0.054 | 1.002 | 1,547.745 |

***Table B9:***

*RQ2: Previous Turn Acknowledgment Parameter Estimates*

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| Negotiation | speaker_typetarget | Fixed | 0.101 | 0.082 | -0.063 | 0.260 | 1.001 | 6,322.464 |
| | speaker_typepartner | Fixed | 0.054 | 0.068 | -0.077 | 0.199 | 1.000 | 6,534.535 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.074 | 0.127 | -0.172 | 0.323 | 1.001 | 3,739.755 |
| | sigma | Residual | 0.036 | 0.033 | 0.002 | 0.122 | 1.004 | 829.394 |
| | sd(Intercept) | Random | 0.044 | 0.041 | 0.001 | 0.153 | 1.002 | 7,562.883 |
| | sd(Intercept) | Random | 0.044 | 0.041 | 0.001 | 0.150 | 1.001 | 4,506.563 |
| | sd(Intercept) | Random | 0.034 | 0.033 | 0.001 | 0.123 | 1.003 | 2,000.790 |
| Round Robin (Video Mediated) | speaker_typetarget | Fixed | 0.194 | 0.098 | -0.012 | 0.376 | 1.010 | 310.751 |
| | speaker_typepartner | Fixed | 0.082 | 0.093 | -0.108 | 0.269 | 1.047 | 852.751 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.105 | 0.115 | -0.113 | 0.347 | 1.022 | 700.745 |
| | sigma | Residual | 0.014 | 0.020 | 0.001 | 0.070 | 1.067 | 53.994 |
| | sd(Intercept) | Random | 0.086 | 0.075 | 0.002 | 0.271 | 1.022 | 168.542 |
| | sd(Intercept) | Random | 0.107 | 0.074 | 0.005 | 0.284 | 1.038 | 395.725 |
| | sd(Intercept) | Random | 0.075 | 0.070 | 0.001 | 0.253 | 1.082 | 34.435 |

***Table B10:***

*RQ2: Previous Turn Agreement Parameter Estimates*

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| Middle Eastern Men | speaker_typetarget | Fixed | 0.102 | 0.125 | -0.134 | 0.356 | 1.013 | 456.611 |
| | speaker_typepartner | Fixed | 0.057 | 0.129 | -0.193 | 0.318 | 1.004 | 644.110 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.145 | 0.139 | -0.125 | 0.420 | 1.005 | 645.021 |
| | sigma | Residual | 0.041 | 0.000 | 0.041 | 0.042 | 1.005 | 760.467 |
| | sd(Intercept) | Random | 0.001 | 0.000 | 0.000 | 0.001 | 1.002 | 732.091 |
| | sd(Intercept) | Random | 0.123 | 0.037 | 0.072 | 0.213 | 1.004 | 295.704 |
| | sd(Intercept) | Random | 0.001 | 0.000 | 0.000 | 0.002 | 1.013 | 674.559 |
| Negotiation | speaker_typetarget | Fixed | 0.115 | 0.015 | 0.086 | 0.144 | 1.000 | 6,345.968 |
| | speaker_typepartner | Fixed | 0.111 | 0.015 | 0.084 | 0.139 | 1.000 | 6,379.732 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.058 | 0.019 | 0.019 | 0.096 | 1.000 | 7,627.004 |
| | sigma | Residual | 0.111 | 0.004 | 0.104 | 0.119 | 1.000 | 6,589.792 |
| | sd(Intercept) | Random | 0.007 | 0.005 | 0.000 | 0.020 | 1.000 | 6,192.318 |
| | sd(Intercept) | Random | 0.007 | 0.006 | 0.000 | 0.021 | 1.000 | 5,965.108 |
| | sd(Intercept) | Random | 0.046 | 0.008 | 0.031 | 0.061 | 1.000 | 5,053.172 |
| Round Robin (Video Mediated) | speaker_typetarget | Fixed | 0.134 | 0.008 | 0.124 | 0.155 | 1.001 | 4,856.377 |
| | speaker_typepartner | Fixed | 0.135 | 0.004 | 0.126 | 0.143 | 1.000 | 6,425.918 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.010 | 0.006 | -0.002 | 0.021 | 1.000 | 6,166.325 |
| | sigma | Residual | 0.066 | 0.001 | 0.065 | 0.067 | 1.000 | 7,146.540 |
| | sd(Intercept) | Random | 0.005 | 0.003 | 0.000 | 0.010 | 1.003 | 1,676.132 |
| | sd(Intercept) | Random | 0.017 | 0.002 | 0.013 | 0.020 | 1.000 | 4,300.459 |
| | sd(Intercept) | Random | 0.030 | 0.001 | 0.027 | 0.033 | 1.001 | 4,443.998 |

***Table B11:***

*RQ2: Previous Turn Disagreement Parameter Estimates*

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| Middle Eastern Men | speaker_typetarget | Fixed | 0.136 | 0.080 | -0.022 | 0.292 | 1.000 | 7,432.520 |
| | speaker_typepartner | Fixed | 0.209 | 0.088 | 0.033 | 0.381 | 1.000 | 8,045.503 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.047 | 0.110 | -0.164 | 0.267 | 1.000 | 8,761.892 |
| | sigma | Residual | 0.096 | 0.001 | 0.094 | 0.098 | 1.000 | 7,698.800 |
| | sd(Intercept) | Random | 0.001 | 0.001 | 0.000 | 0.003 | 1.001 | 8,219.291 |
| | sd(Intercept) | Random | 0.119 | 0.047 | 0.060 | 0.238 | 1.000 | 6,900.169 |
| | sd(Intercept) | Random | 0.001 | 0.001 | 0.000 | 0.004 | 1.000 | 7,539.428 |
| Negotiation | speaker_typetarget | Fixed | 0.101 | 0.028 | 0.046 | 0.156 | 1.000 | 8,834.450 |
| | speaker_typepartner | Fixed | 0.093 | 0.027 | 0.042 | 0.146 | 1.000 | 8,474.233 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.158 | 0.036 | 0.089 | 0.229 | 1.000 | 8,647.758 |
| | sigma | Residual | 0.139 | 0.007 | 0.126 | 0.153 | 1.000 | 7,102.283 |
| | sd(Intercept) | Random | 0.034 | 0.022 | 0.002 | 0.080 | 1.001 | 3,267.270 |
| | sd(Intercept) | Random | 0.035 | 0.022 | 0.002 | 0.083 | 1.000 | 3,134.718 |
| | sd(Intercept) | Random | 0.025 | 0.018 | 0.001 | 0.066 | 1.000 | 3,686.287 |
| Round Robin (Video Mediated) | speaker_typetarget | Fixed | 0.122 | 0.007 | 0.109 | 0.134 | 1.001 | 7,380.521 |
| | speaker_typepartner | Fixed | 0.123 | 0.007 | 0.111 | 0.135 | 1.000 | 7,496.306 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.007 | 0.008 | -0.008 | 0.022 | 1.000 | 8,617.820 |
| | sigma | Residual | 0.072 | 0.001 | 0.070 | 0.074 | 1.000 | 7,146.909 |
| | sd(Intercept) | Random | 0.006 | 0.004 | 0.000 | 0.015 | 1.000 | 2,014.759 |
| | sd(Intercept) | Random | 0.008 | 0.004 | 0.001 | 0.017 | 1.001 | 1,814.536 |
| | sd(Intercept) | Random | 0.039 | 0.003 | 0.034 | 0.044 | 1.000 | 4,634.221 |

***Table B12:***

*RQ2: Previous Turn Functional Parameter Estimates*

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| Middle Eastern Men | speaker_typetarget | Fixed | 0.222 | 0.147 | -0.067 | 0.506 | 1.014 | 198.372 |
| | speaker_typepartner | Fixed | 0.188 | 0.109 | -0.028 | 0.389 | 1.031 | 208.648 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.273 | 0.143 | -0.021 | 0.547 | 1.016 | 253.601 |
| | sigma | Residual | 0.052 | 0.001 | 0.051 | 0.053 | 1.017 | 329.500 |
| | sd(Intercept) | Random | 0.001 | 0.000 | 0.000 | 0.002 | 1.019 | 345.140 |
| | sd(Intercept) | Random | 0.163 | 0.060 | 0.090 | 0.325 | 1.058 | 74.373 |
| | sd(Intercept) | Random | 0.001 | 0.001 | 0.000 | 0.002 | 1.017 | 270.850 |
| Negotiation | speaker_typetarget | Fixed | 0.172 | 0.038 | 0.099 | 0.249 | 1.000 | 8,861.868 |
| | speaker_typepartner | Fixed | 0.129 | 0.039 | 0.048 | 0.202 | 1.000 | 8,414.755 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.073 | 0.054 | -0.029 | 0.183 | 1.000 | 8,301.233 |
| | sigma | Residual | 0.135 | 0.011 | 0.115 | 0.158 | 1.001 | 7,001.139 |
| | sd(Intercept) | Random | 0.024 | 0.019 | 0.001 | 0.071 | 1.000 | 6,971.358 |
| | sd(Intercept) | Random | 0.024 | 0.019 | 0.001 | 0.072 | 1.000 | 7,211.704 |
| | sd(Intercept) | Random | 0.032 | 0.022 | 0.001 | 0.081 | 1.000 | 5,464.363 |
| Round Robin (Video Mediated) | speaker_typetarget | Fixed | 0.128 | 0.013 | 0.104 | 0.153 | 1.000 | 7,087.586 |
| | speaker_typepartner | Fixed | 0.125 | 0.013 | 0.099 | 0.150 | 1.000 | 7,558.379 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.016 | 0.018 | -0.019 | 0.051 | 1.000 | 7,831.010 |
| | sigma | Residual | 0.074 | 0.003 | 0.068 | 0.079 | 1.000 | 7,104.509 |
| | sd(Intercept) | Random | 0.016 | 0.010 | 0.001 | 0.036 | 1.002 | 1,833.450 |
| | sd(Intercept) | Random | 0.013 | 0.009 | 0.001 | 0.032 | 1.002 | 2,272.375 |
| | sd(Intercept) | Random | 0.047 | 0.007 | 0.033 | 0.059 | 1.001 | 2,974.745 |

***Table B13:***

*RQ2: Previous Turn Inform Parameter Estimates*

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| Negotiation | speaker_typetarget | Fixed | 0.131 | 0.023 | 0.088 | 0.178 | 1.000 | 8,127.184 |
| | speaker_typepartner | Fixed | 0.119 | 0.015 | 0.090 | 0.149 | 1.000 | 8,351.902 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.127 | 0.020 | 0.087 | 0.167 | 1.001 | 8,762.589 |
| | sigma | Residual | 0.147 | 0.003 | 0.141 | 0.153 | 1.000 | 6,707.223 |
| | sd(Intercept) | Random | 0.023 | 0.013 | 0.001 | 0.048 | 1.001 | 2,128.268 |
| | sd(Intercept) | Random | 0.024 | 0.013 | 0.001 | 0.049 | 1.001 | 2,244.921 |
| | sd(Intercept) | Random | 0.042 | 0.008 | 0.028 | 0.057 | 1.001 | 5,155.681 |

***Table B14:***

*RQ2: Previous Turn Question Parameter Estimates*

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| Negotiation | speaker_typetarget | Fixed | 0.138 | 0.025 | 0.089 | 0.185 | 1.000 | 7,022.853 |
| | speaker_typepartner | Fixed | 0.139 | 0.022 | 0.094 | 0.180 | 1.000 | 6,681.915 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.061 | 0.031 | 0.002 | 0.122 | 1.000 | 7,153.419 |
| | sigma | Residual | 0.141 | 0.006 | 0.130 | 0.153 | 1.000 | 5,424.021 |
| | sd(Intercept) | Random | 0.014 | 0.010 | 0.001 | 0.038 | 1.000 | 5,322.941 |
| | sd(Intercept) | Random | 0.014 | 0.010 | 0.000 | 0.037 | 1.001 | 5,439.111 |
| | sd(Intercept) | Random | 0.039 | 0.017 | 0.004 | 0.070 | 1.002 | 2,475.349 |
| Round Robin (Video Mediated) | speaker_typetarget | Fixed | 0.137 | 0.009 | 0.121 | 0.154 | 1.001 | 5,160.078 |
| | speaker_typepartner | Fixed | 0.130 | 0.012 | 0.111 | 0.151 | 1.001 | 3,976.809 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.011 | 0.011 | -0.008 | 0.036 | 1.001 | 5,084.355 |
| | sigma | Residual | 0.083 | 0.001 | 0.081 | 0.085 | 1.000 | 6,775.146 |
| | sd(Intercept) | Random | 0.004 | 0.003 | 0.000 | 0.009 | 1.002 | 2,318.776 |
| | sd(Intercept) | Random | 0.010 | 0.003 | 0.003 | 0.016 | 1.005 | 1,443.884 |
| | sd(Intercept) | Random | 0.035 | 0.002 | 0.032 | 0.039 | 1.002 | 3,474.029 |

***Table B15:***

*RQ2: Previous Turn Suggestion Parameter Estimates*

| Study | Parameter | Class | Estimate | SE | l-95% CI | u-95% CI | Rhat | ESS |
|---|---|---|---|---|---|---|---|---|
| Negotiation | speaker_typetarget | Fixed | 0.162 | 0.077 | 0.006 | 0.308 | 1.000 | 8,961.423 |
| | speaker_typepartner | Fixed | 0.149 | 0.069 | 0.010 | 0.284 | 1.000 | 9,204.754 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.040 | 0.095 | -0.137 | 0.236 | 1.000 | 8,919.821 |
| | sigma | Residual | 0.101 | 0.020 | 0.070 | 0.147 | 1.000 | 7,682.415 |
| | sd(Intercept) | Random | 0.049 | 0.042 | 0.002 | 0.154 | 1.000 | 7,834.798 |
| | sd(Intercept) | Random | 0.049 | 0.042 | 0.002 | 0.152 | 1.000 | 7,792.268 |
| | sd(Intercept) | Random | 0.045 | 0.038 | 0.001 | 0.140 | 1.000 | 7,834.806 |
| Round Robin (Video Mediated) | speaker_typetarget | Fixed | 0.133 | 0.029 | 0.077 | 0.191 | 1.000 | 7,889.828 |
| | speaker_typepartner | Fixed | 0.134 | 0.026 | 0.080 | 0.186 | 1.000 | 8,068.013 |
| | Ipercent_of_chapter_at_utterance_startE2 | Fixed | 0.003 | 0.036 | -0.067 | 0.074 | 1.000 | 7,211.342 |
| | sigma | Residual | 0.076 | 0.007 | 0.063 | 0.090 | 1.000 | 6,400.748 |
| | sd(Intercept) | Random | 0.020 | 0.014 | 0.001 | 0.053 | 1.001 | 4,137.789 |
| | sd(Intercept) | Random | 0.024 | 0.016 | 0.001 | 0.059 | 1.002 | 3,426.171 |
| | sd(Intercept) | Random | 0.021 | 0.015 | 0.001 | 0.055 | 1.001 | 3,741.567 |

## REFERENCES CITED


Anikin, A., Bååth, R., & Persson, T. (2018). Human non-linguistic vocal repertoire: Call types and their meaning. *Journal of Nonverbal Behavior*, *42*(1), 53–80. https://doi.org/10.1007/s10919-017-0267-y

Atias, D., Aviezer, H., Ochsner, K., & Gilead, M. (2021). Empathic accuracy: Lessons from the perception of contextualized real-life emotional expressions. In M. Gilead & K. Ochsner (Eds.), *The neural basis of mentalizing* (pp. 171–188). Springer International Publishing. https://doi.org/10.1007/978-3-030-51890-5_9

Atias, D., Todorov, A., Liraz, S., Eidinger, A., Dror, I., Maymon, Y., & Aviezer, H. (2019). Loud and unclear: Intense real-life vocalizations during affective situations are perceptually ambiguous and contextually malleable. *Journal of Experimental Psychology: General*, *148*(10), 1842–1848. https://doi.org/10.1037/xge0000535

Austin, J. L. (1962). *How to do things with words*. Oxford University Press.

Aviezer, H., Ensenberg, N., & Hassin, R. R. (2017). The inherently contextualized nature of facial emotion perception. *Current Opinion in Psychology*, *17*, 47–54. https://doi.org/10.1016/j.copsyc.2017.06.006

Bales, R. F. (1950). *Interaction process analysis; a method for the study of small groups* (pp. xi, 203). Addison-Wesley.

Bavelas, J. B., Coates, L., & Johnson, T. (2000). Listeners as co-narrators. *Journal of Personality and Social Psychology*, *79*(6), 941–952. https://doi.org/10.1037/0022-3514.79.6.941

Belias, M., Rovers, M. M., Hoogland, J., Reitsma, J. B., Debray, T. P. A., & IntHout, J. (2022). Predicting personalised absolute treatment effects in individual participant data meta-

analysis: An introduction to splines. *Research Synthesis Methods*, *13*(2), 255–283. https://doi.org/10.1002/jrsm.1546

Berlamont, L., Ceulemans, E., Carlier, C., Verhofstadt, L., Ickes, W., Hinnekens, C., & Sels, L. (2024). Couple similarity in empathic accuracy and relationship well-being. *Journal of Social and Personal Relationships*, *41*(6), 1301–1323. https://doi.org/10.1177/02654075231206412

Berlamont, L., Hodges, S., Sels, L., Ceulemans, E., Ickes, W., Hinnekens, C., & Verhofstadt, L. (2023). Motivation and empathic accuracy during conflict interactions in couples: It's complicated! *Motivation and Emotion*, *47*(2), 208–228. https://doi.org/10.1007/s11031-022-09982-x

Bernieri, F. J., Gillis, J. S., Davis, J. M., & Grahe, J. E. (1996). Dyad rapport and the accuracy of its judgment across situations: A lens model analysis. *Journal of Personality and Social Psychology*, *71*(1), 110–129. https://doi.org/10.1037/0022-3514.71.1.110

Bernieri, F. J., & Gillis, J. S. (2001). Judging rapport: Employing Brunswik's lens model to study interpersonal sensitivity. In J. A. Hall & F. J. Bernieri (Eds.), *Interpersonal sensitivity: Theory and measurement* (pp. 67–88). Lawrence Erlbaum Associates Publishers.

Bhowmik, C. V., Back, M. D., Nestler, S., & Schrader, F.-W. (2025). Appearing smart, confident and motivated: A lens model approach to judgment accuracy in an educational setting. *Social Psychology of Education*, *28*(1), 105–138. https://doi.org/10.1007/s11218-025-10057-1

Biron, T., Baum, D., Freche, D., Matalon, N., Ehrmann, N., Weinreb, E., Biron, D., & Moses, E. (2021). Automatic detection of prosodic boundaries in spontaneous speech. *PLOS One*, *16*(5), e0250969. https://doi.org/10.1371/journal.pone.0250969

Bodie, G. D., Jones, S. M., Brinberg, M., Joyer, A. M., Solomon, D. H., & Ram, N. (2021). Discovering the fabric of supportive conversations: A typology of speaking turns and their contingencies. *Journal of Language and Social Psychology*, *40*(2), 214–237. https://doi.org/10.1177/0261927X20953604

Breil, S. M., Osterholz, S., Nestler, S., & Back, M. D. (2021). Contributions of nonverbal cues to the accurate judgment of personality traits. In T. D. Letzring & J. S. Spain (Eds.), *The Oxford handbook of accurate personality judgment* (1st ed., pp. 194–218). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190912529.013.13

Bristow, D. N., Mowen, J. C., & Krieger, R. H. (2015). The quality lens model: A marketing tool for improving channel relationships. In E. J. Wilson & W. C. Black (Eds.), *Proceedings of the 1994 Academy of Marketing Science (AMS) annual conference* (pp. 397–401). Springer International Publishing. https://doi.org/10.1007/978-3-319-13162-7_108

Briton, N. J., & Hall, J. A. (1995). Beliefs about female and male nonverbal communication. *Sex Roles*, *32*(1–2), 79–90. https://doi.org/10.1007/BF01544758

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*(3), 193–217. https://doi.org/10.1037/h0047470

Brunswik, E. (1956). *Perception and the representative design of psychological experiments, 2nd ed*. University of California Press.

Bryant, G. A., & Fox Tree, J. E. (2005). Is there an ironic tone of voice? *Language and Speech*, *48*(3), 257–277. https://doi.org/10.1177/00238309050480030101

Buck, R., Powers, S. R., & Hull, K. S. (2017). Measuring emotional and cognitive empathy using dynamic, naturalistic, and spontaneous emotion displays. *Emotion*, *17*(7), 1120–1136. https://doi.org/10.1037/emo0000285

Burke, D. L., Ensor, J., & Riley, R. D. (2017). Meta-analysis using individual participant data: One-stage and two-stage approaches, and why they may differ. *Statistics in Medicine*, *36*(5), 855–875. https://doi.org/10.1002/sim.7141

Bürkner, P.-C. (2017). **brms**: An *R* package for Bayesian multilevel models using *Stan*. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Burtsev, M., Seliverstov, A., Airapetyan, R., Arkhipov, M., Baymurzina, D., Bushkov, N., Gureenkova, O., Khakhulin, T., Kuratov, Y., Kuznetsov, D., Litinsky, A., Logacheva, V., Lymar, A., Malykh, V., Petrov, M., Polulyakh, V., Pugachev, L., Sorokin, A., Vikhreva, M., & Zaynutdinov, M. (2018). DeepPavlov: Open-source library for dialogue systems. *Proceedings of ACL 2018, System Demonstrations*, 122–127. https://doi.org/10.18653/v1/P18-4021

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). *Stan*: A probabilistic programming language. *Journal of Statistical Software*, *76*(1). https://doi.org/10.18637/jss.v076.i01

Chen, Z., & Whitney, D. (2019). Tracking the affective state of unseen persons. *Proceedings of the National Academy of Sciences*, *116*(15), 7559–7564. https://doi.org/10.1073/pnas.1812250116

Chinsky, J. M., & Rappaport, J. (1970). Brief critique of the meaning and reliability of "accurate empathy" ratings. *Psychological Bulletin*, *73*(5), 379–382. https://doi.org/10.1037/h0029224

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on Socially Shared Cognition* (pp. 127–149). American Psychological Association. https://doi.org/10.1037/10096-006

Core, M. G., & Allen, J. F. (1997). Coding dialogs with the DAMSL annotation scheme. *Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, 28–35.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*(1), 113–126. https://doi.org/10.1037/0022-3514.44.1.113

Davis, M. H., Konrath, S., Briethaupt, F., Cameron, C. D., Decety, J., de Waal, F., Eisenberg, N., Hall, J. A., Hodges, S. D., Keen, S., Koopmann-Holm, B., Maibom, H., Preston, S. D., Vorauer, J. D., & Zaki, J. (2026). Mapping the empathy universe: Toward a common lexicon of empathy-related constructs. *Perspectives on Psychological Science*.

Debray, T. P. A., Moons, K. G. M., Van Valkenhoef, G., Efthimiou, O., Hummel, N., Groenwold, R. H. H., & Reitsma, J. B. (2015). Get real in individual participant data (IPD) meta-analysis: A review of the methodology. *Research Synthesis Methods*, *6*(4), 293–309. https://doi.org/10.1002/jrsm.1160

Degen, J. (2023). The rational speech act framework. *Annual Review of Linguistics*, *9*(Volume 9, 2023), 519–540. https://doi.org/10.1146/annurev-linguistics-031220-010811

DesJardins, N. M. L., & Hodges, S. D. (2015). Reading between the lies: Empathic accuracy and deception detection. *Social Psychological and Personality Science*, *6*(7), 781–787. https://doi.org/10.1177/1948550615585829

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dogan, A., & Birant, D. (2019). A weighted majority voting ensemble approach for classification. *2019 4th International Conference on Computer Science and Engineering (UBMK)*, 1–6. https://doi.org/10.1109/UBMK.2019.8907028

Dougherty, T. W., Ebert, R. J., & Callender, J. C. (1986). Policy capturing in the employment interview. *Journal of Applied Psychology*, *71*(1), 9–15. https://doi.org/10.1037/0021-9010.71.1.9

Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & Baribault, B. (2018). How to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin & Review*, *25*(1), 219–234. https://doi.org/10.3758/s13423-017-1317-5

Fernandez-Duque, D., Hodges, S. D., Baird, J. A., & Black, S. E. (2010). Empathy in frontotemporal dementia and Alzheimer's disease. *Journal of Clinical and Experimental Neuropsychology*, *32*(3), 289–298. https://doi.org/10.1080/13803390903002191

Frohmann, M., Sterner, I., Vulić, I., Minixhofer, B., & Schedl, M. (2024). *Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation* (arXiv:2406.16678). arXiv. https://doi.org/10.48550/ARXIV.2406.16678

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*(4), 652–670. https://doi.org/10.1037/0033-295X.102.4.652

Funder, D. C., & Colvin, C. R. (1988). Friends and strangers: Acquaintanceship, agreement, and the accuracy of personality judgment. *Journal of Personality and Social Psychology*, *55*(1), 149–158. https://doi.org/10.1037/0022-3514.55.1.149

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3). https://doi.org/10.1214/06-BA117A

Gelman, A., Simpson, D., & Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy*, *19*(10), 555. https://doi.org/10.3390/e19100555

Gesn, P. R., & Ickes, W. (1999). The development of meaning contexts for empathic accuracy: Channel and sequence effects. *Journal of Personality and Social Psychology*, *77*(4), 746–761. https://doi.org/10.1037/0022-3514.77.4.746

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal*, *31*(1), 104–137.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Hall, J. A. (1978). Gender effects in decoding nonverbal cues. *Psychological Bulletin*, *85*(4), 845–857. https://doi.org/10.1037/0033-2909.85.4.845

Hall, J. A., Gunnery, S. D., & Schlegel, K. (2025). Gender and accuracy in decoding affect cues: A meta-analysis. *Journal of Intelligence*, *13*(3), 38–67. https://doi.org/10.3390/jintelligence13030038

Hall, J. A., & Schmid Mast, M. (2007). Sources of accuracy in the empathic accuracy paradigm. *Emotion*, *7*(2), 438–446. https://doi.org/10.1037/1528-3542.7.2.438

Harrell, F. E. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis*. Springer International Publishing. https://doi.org/10.1007/978-3-319-19425-7

Hartwig, M., & Bond, C. F. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin*, *137*(4), 643–659. https://doi.org/10.1037/a0023589

Hartwig, M., & Bond, C. F. (2014). Lie detection from multiple cues: A meta-analysis. *Applied Cognitive Psychology*, *28*(5), 661–676. https://doi.org/10.1002/acp.3052

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

He, P., Liu, X., Gao, J., & Chen, W. (2021). *DeBERTa: Decoding-enhanced BERT with disentangled attention* (arXiv:2006.03654). arXiv. https://doi.org/10.48550/arXiv.2006.03654

Hinnekens, C., Ickes, W., Berlamont, L., & Verhofstadt, L. (2021). Empathic accuracy: Empirical overview and clinical applications. In M. Gilead & K. Ochsner (Eds.), *The neural basis of mentalizing* (pp. 149–170). Springer International Publishing. https://doi.org/10.1007/978-3-030-51890-5_8

Hodges, S. D., & Kezer, M. (2021). It is hard to read minds without words: Cues to use to achieve empathic accuracy. *Journal of Intelligence*, *9*(2), 27–49. https://doi.org/10.3390/jintelligence9020027

Hodges, S. D., Kezer, M., Hall, J. A., & Vorauer, J. D. (2024). Exploring actual and presumed links between accurately inferring contents of other people's minds and prosocial outcomes. *Journal of Intelligence*, *12*(2), 13–30. https://doi.org/10.3390/jintelligence12020013

Hodges, S. D., Kiel, K. J., Kramer, A. D. I., Veach, D., & Villanueva, B. R. (2010). Giving birth to empathy: The effects of similar experience on empathic accuracy, empathic concern, and perceived empathy. *Personality and Social Psychology Bulletin*, *36*(3), 398–409. https://doi.org/10.1177/0146167209350326

Hodges, S. D., Laurent, S. M., & Lewis, K. L. (2011). Specially motivated, feminine, or just female: Do women have an empathic accuracy advantage. In J. L. Smith, W. Ickes, J. A. Hall, & S. D. Hodges (Eds.), *Managing interpersonal sensitivity: Knowing when-and when not-to understand others* (pp. 59–74). Nova Science Publishers.

Hodges, S. D., Lewis, K. L., & Ickes, W. (2015). The matter of other minds: Empathic accuracy and the factors that influence it. In M. Mikulincer, P. R. Shaver, J. A. Simpson, & J. F. Dovidio (Eds.), *APA handbook of personality and social psychology, volume 3: Interpersonal relations.* (pp. 319–348). American Psychological Association. https://doi.org/10.1037/14344-012

Hodges, S. D., Lewis, K. L., Wise, A. A. W., Bogart, K., Bruun, M., Christian, C., Denning, K. R., & Jacobs, E. (2022). *Sex differences in perceptions of feedback in STEM fields* [Dataset]. University of Oregon.

Howington, D. E. (2016). *What does accuracy get you? Empathic accuracy during a negotiation* [Dissertation, University of Oregon]. UO Scholars' Bank Psychology Dissertation and Theses. https://hdl.handle.net/1794/22298

Huang, K., Yeomans, M., Brooks, A. W., Minson, J., & Gino, F. (2017). It doesn't hurt to ask: Question-asking increases liking. *Journal of Personality and Social Psychology*, *113*(3), 430–452. https://doi.org/10.1037/pspi0000097

Ickes, W. (1982). A basic paradigm for the study of personality, roles, and social behavior. In W. Ickes & E. S. Knowles (Eds.), *Personality, roles, and social behavior* (pp. 305–341). Springer New York. https://doi.org/10.1007/978-1-4613-9469-3_11

Ickes, W. (1993). Empathic accuracy. *Journal of Personality*, *61*(4), 587–610. https://doi.org/10.1111/j.1467-6494.1993.tb00783.x

Ickes, W. (2001). Measuring empathic accuracy. In J. A. Hall & F. J. Bernieri (Eds.), *Interpersonal sensitivity: Theory and measurement* (pp. 219–241). Lawrence Erlbaum Associates.

Ickes, W. (2011). Everyday mind reading is driven by motives and goals. *Psychological Inquiry*, *22*(3), 200–206. https://doi.org/10.1080/1047840X.2011.561133

Ickes, W. (2016). Empathic accuracy: Judging thoughts and feelings. In J. A. Hall, M. Schmid Mast, & T. V. West (Eds.), *The social psychology of perceiving others accurately* (pp. 52–70). Cambridge University Press.

Ickes, W., Gesn, P. R., & Graham, T. (2000). Gender differences in empathic accuracy: Differential ability or differential motivation? *Personal Relationships*, *7*(1), 95–109. https://doi.org/10.1111/j.1475-6811.2000.tb00006.x

Ickes, W., Stinson, L., Bissonnette, V., & Garcia, S. (1990). Naturalistic social cognition: Empathic accuracy in mixed-sex dyads. *Journal of Personality and Social Psychology*, *59*(4), 730–742. https://doi.org/10.1037/0022-3514.59.4.730

Jiang, Z., Yu, W., Zhou, D., Chen, Y., Feng, J., & Yan, S. (2020). *ConvBERT: Improving BERT with span-based dynamic convolution* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2008.02496

Jourard, S. M. (1971). *Self-disclosure: An experimental analysis of the transparent self*. Wiley.

Jurafsky, D., Shriberg, E., & Biasca, D. (1997). *Switchboard SWBDDAMSL ShallowDiscourse-Function Annotation Coders Manual*. ICS Technical Report. http://stripe.colorado.edu/~jurafsky/manual.august1.html

Kagan, N. (1977). *Interpersonal process recall*. Michigan State University Press.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin*, *134*(3), 404–426. https://doi.org/10.1037/0033-2909.134.3.404

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kezer, M. (doctoral dissertation, in prep). [Doctoral dissertation]. University of Oregon.

Kezer, M., Schroeder, Z. J., & Hodges, S. D. (2025). *Computational modeling of thought-feeling accuracy using sentence embeddings*. PsyArXiv. https://doi.org/10.31234/osf.io/t93kh_v1

Kiesler, D. J., Mathieu, P. L., & Klein, M. H. (1967). Patient experiencing level and interaction-chronograph variables in therapy interview segments. *Journal of Consulting Psychology*, *31*(2), 224–224. https://doi.org/10.1037/h0024434

Kilpatrick, S. D., Bissonnette, V. L., & Rusbult, C. E. (2002). Empathic accuracy and accommodative behavior among newly married couples. *Personal Relationships*, *9*(4), 369–393. https://doi.org/10.1111/1475-6811.09402

Klein, K. J. K., & Hodges, S. D. (2001). Gender differences, motivation, and empathic accuracy: When it pays to understand. *Personality and Social Psychology Bulletin*, *27*(6), 720–730. https://doi.org/10.1177/0146167201276007

Kraus, M. W. (2017). Voice-only communication enhances empathic accuracy. *American Psychologist*, *72*(7), 644–654. https://doi.org/10.1037/amp0000147

Kruschke, J. K. (2021). Bayesian analysis reporting guidelines. *Nature Human Behaviour*, *5*(10), 1282–1291. https://doi.org/10.1038/s41562-021-01177-7

Küfner, A. C. P., Back, M. D., Nestler, S., & Egloff, B. (2010). Tell me a story and I will tell you who you are! Lens model analyses of personality and creative writing. *Journal of Research in Personality*, *44*(4), 427–435. https://doi.org/10.1016/j.jrp.2010.05.003

Kunzmann, U., Wieck, C., & Dietzel, C. (2018). Empathic accuracy: Age differences from adolescence into middle adulthood. *Cognition and Emotion*, *32*(8), 1611–1624. https://doi.org/10.1080/02699931.2018.1433128

Levenson, R. W., & Gottman, J. M. (1983). Marital interaction: Physiological linkage and affective exchange. *Journal of Personality and Social Psychology*, *45*(3), 587–597. https://doi.org/10.1037/0022-3514.45.3.587

Levenson, R. W., & Ruef, A. M. (1992). Empathy: A physiological substrate. *Journal of Personality and Social Psychology*, *63*(2), 234–246. https://doi.org/10.1037/0022-3514.63.2.234

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

Lewis, K. L. (2014). *Searching for the open book: Exploring predictors of target readability in interpersonal accuracy* [Dissertation, University of Oregon]. UO Scholars' Bank Psychology Dissertation and Theses. https://hdl.handle.net/1794/18374

Lewis, K. L., Hodges, S. D., Laurent, S. M., Srivastava, S., & Biancarosa, G. (2012). Reading between the minds: The use of stereotypes in empathic accuracy. *Psychological Science*, *23*(9), 1040–1046. https://doi.org/10.1177/0956797612439719

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., & Niu, S. (2017). *DailyDialog: A manually labelled multi-turn dialogue dataset* (arXiv:1710.03957). arXiv. https://doi.org/10.48550/arXiv.1710.03957

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach* (arXiv:1907.11692). arXiv. https://doi.org/10.48550/ARXIV.1907.11692

Lucas-Molina, B., Pérez-Albéniz, A., Ortuño-Sierra, J., & Fonseca-Pedrero, E. (2017). Dimensional structure and measurement invariance of the interpersonal reactivity index (IRI) across gender. *Psicothema*, *4*(29), 590–595. https://doi.org/10.7334/psicothema2017.19

Lunn, D., Barrett, J., Sweeting, M., & Thompson, S. (2013). Fully Bayesian hierarchical modelling in two stages, with application to meta-analysis. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, *62*(4), 551–572. https://doi.org/10.1111/rssc.12007

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86

Ma-Kellams, C., & Blascovich, J. (2013). The ironic effect of financial incentive on empathic accuracy. *Journal of Experimental Social Psychology*, *49*(1), 65–71. https://doi.org/10.1016/j.jesp.2012.08.014

Marangoni, C., Garcia, S., Ickes, W., & Teng, G. (1995). Empathic accuracy in a clinically relevant setting. *Journal of Personality and Social Psychology*, *68*(5), 854–869. https://doi.org/10.1037/0022-3514.68.5.854

Martingano, A. J., Konrath, S., Blanch-Hartigan, D., Davis, M. H., Hall, J. A., Ruben, M. A., Sanders, J. J., & Schwartz, R. (2025). Seven empathy myths: Correcting misconceptions about a complex construct. *Social and Personality Psychology Compass*, *19*(10), e70093. https://doi.org/10.1111/spc3.70093

Matsui, T., Nakamura, T., Utsumi, A., Sasaki, A. T., Koike, T., Yoshida, Y., Harada, T., Tanabe, H. C., & Sadato, N. (2016). The role of prosody and context in sarcasm comprehension: Behavioral and fMRI evidence. *Neuropsychologia*, *87*, 74–84. https://doi.org/10.1016/j.neuropsychologia.2016.04.031

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and stan* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429029608

McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, *28*(2), 295–314. https://doi.org/10.1007/s10648-014-9287-x

Oehlert, G. W. (1992). A note on the delta method. *American Statistician*, *46*(1), 27–29. https://doi.org/10.1080/00031305.1992.10475842

Osterholz, S., Breil, S. M., Nestler, S., & Back, M. D. (2019). Lens and dual lens models. In T. D. Letzring & J. S. Spain (Eds.), *The Oxford handbook of accurate personality judgment*

(1st ed., pp. 45–60). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190912529.013.4

Paananen, T., Piironen, J., Bürkner, P.-C., & Vehtari, A. (2021). Implicitly adaptive importance sampling. *Statistics and Computing*, *31*(2), 16. https://doi.org/10.1007/s11222-020-09982-2

Patterson, M. L., Fridlund, A. J., & Crivelli, C. (2023). Four misconceptions about nonverbal communication. *Perspectives on Psychological Science*, *18*(6), 1388–1411. https://doi.org/10.1177/17456916221148142

Polson, N. G., & Scott, J. G. (2012). On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, *7*(4). https://doi.org/10.1214/12-BA730

Prahallad, L., & Mamidi, R. (2025). *Analyzing biases in political dialogue: Tagging U.S. presidential debates with an extended DAMSL framework* (arXiv:2505.19515). arXiv. https://doi.org/10.48550/ARXIV.2505.19515

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust speech recognition via large-scale weak supervision* (arXiv:2212.04356). arXiv. https://doi.org/10.48550/ARXIV.2212.04356

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2023). *Exploring the limits of transfer learning with a unified text-to-text transformer* (arXiv:1910.10683). arXiv. https://doi.org/10.48550/arXiv.1910.10683

Rappaport, J., & Chinsky, J. M. (1972). Accurate empathy: Confusion of a construct. *Psychological Bulletin*, *77*(6), 400–404. https://doi.org/10.1037/h0032716

Rasmussen, C. E., & Williams, C. K. I. (2008). *Gaussian processes for machine learning* (3. print). MIT Press.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3980–3990. https://doi.org/10.18653/v1/D19-1410

Reis, H. T., Lemay, E. P., & Finkenauer, C. (2017). Toward understanding understanding: The importance of feeling understood in relationships. *Social and Personality Psychology Compass*, *11*(3), e12308. https://doi.org/10.1111/spc3.12308

Riley, R. D., Lambert, P. C., & Abo-Zaid, G. (2010). Meta-analysis of individual participant data: Rationale, conduct, and reporting. *BMJ*, *340*(feb05 1), c221–c221. https://doi.org/10.1136/bmj.c221

Riley, R. D., Tierney, J. F., & Stewart, L. A. (Eds.). (2021). *Individual participant data meta-analysis: A handbook for healthcare research* (1st ed.). Wiley. https://doi.org/10.1002/9781119333784

Riutort-Mayol, G., Bürkner, P.-C., Andersen, M. R., Solin, A., & Vehtari, A. (2023). Practical Hilbert space approximate Bayesian Gaussian processes for probabilistic programming. *Statistics and Computing*, *33*(1), 17. https://doi.org/10.1007/s11222-022-10167-2

Rogers, C. R. (1957). The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology*, *21*(2), 95–103. https://doi.org/10.1037/h0045357

Rossignac-Milon, M., Bolger, N., Zee, K. S., Boothby, E. J., & Higgins, E. T. (2021). Merged minds: Generalized shared reality in dyadic relationships. *Journal of Personality and Social Psychology*, *120*(4), 882–911. https://doi.org/10.1037/pspi0000266

Röver, C. (2020). Bayesian random-effects meta-analysis using the **bayesmeta** *R* package. *Journal of Statistical Software*, *93*(6). https://doi.org/10.18637/jss.v093.i06

Röver, C., Bender, R., Dias, S., Schmid, C. H., Schmidli, H., Sturtz, S., Weber, S., & Friede, T. (2021). On weakly informative prior distributions for the heterogeneity parameter in Bayesian random-effects meta-analysis. *Research Synthesis Methods*, *12*(4), 448–474. https://doi.org/10.1002/jrsm.1475

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience. *Behavioral and Brain Sciences*, *36*(4), 393–414. https://doi.org/10.1017/S0140525X12000660

Searle, J. R. (1969). *Speech acts*. Cambridge University Press.

Sillars, A. L., & Scott, M. D. (1983). Interpersonal perception between intimates: An integrative review. *Human Communication Research*, *10*(1), 153–176. https://doi.org/10.1111/j.1468-2958.1983.tb00009.x

Simpson, J. A., Ickes, W., & Blackstone, T. (1995). When the head protects the heart: Empathic accuracy in dating relationships. *Journal of Personality and Social Psychology*, *69*(4), 629–641. https://doi.org/10.1037/0022-3514.69.4.629

Solin, A., & Särkkä, S. (2020). Hilbert space methods for reduced-rank Gaussian process regression. *Statistics and Computing*, *30*(2), 419–446. https://doi.org/10.1007/s11222-019-09886-w

Steck, H., Ekanadham, C., & Kallus, N. (2024). Is cosine-similarity of embeddings really about similarity? *Companion Proceedings of the ACM Web Conference 2024*, 887–890. https://doi.org/10.1145/3589335.3651526

Stein, M. L. (1999). *Interpolation of spatial data*. Springer New York. https://doi.org/10.1007/978-1-4612-1494-6

Stiles, W. B. (1992). *Describing talk: A taxonomy of verbal response modes*. Sage.

Stiles, W. B. (2017). Taxonomy of verbal response modes (VRM). In D. L. Worthington & G. D. Bodie (Eds.), *The sourcebook of listening research* (1st ed., pp. 605–611). Wiley. https://doi.org/10.1002/9781119102991.ch69

Stinson, L., & Ickes, W. (1992). Empathic accuracy in the interactions of male friends versus male strangers. *Journal of Personality and Social Psychology*, *62*(5), 787–797. https://doi.org/10.1037/0022-3514.62.5.787

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C. V., & Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, *26*(3), 339–373. https://doi.org/10.1162/089120100561737

Su, H., & Ye, J. (2025). Large language models for automating fine-grained speech act annotation: A critical evaluation of GPT-4o and DeepSeek. *Corpus Pragmatics*, *9*(4), 463–482. https://doi.org/10.1007/s41701-025-00200-w

TeBlunthuis, N., Hase, V., & Chan, C.-H. (2024). Misclassification in automated content analysis causes bias in regression. Can we fix it? Yes we can! *Communication Methods and Measures*, *18*(3), 278–299. https://doi.org/10.1080/19312458.2023.2293713

The Behavior Panel. (2021, March 12). *Meghan & Harry: The WARNING signs* [Video recording]. https://www.youtube.com/watch?v=AYyEx20DiKU

Truax, C. B. (1961). The process of group psychotherapy: Relationships between hypothesized therapeutic conditions and intrapersonal exploration. *Psychological Monographs: General and Applied*, *75*(7), 1–35. https://doi.org/10.1037/h0093771

Truax, C. B. (1972). The meaning and reliability of accurate empathy ratings: A rejoinder. *Psychological Bulletin*, *77*(6), 397–399. https://doi.org/10.1037/h0032718

Truax, C. B., & Carkhuff, R. R. (1967). *Toward effective counseling and psychotherapy: Training and practice* (pp. xiv, 416). Aldine Publishing Co.

Tucker, L. R. (1964). A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch, and by Hammond, Hursch, and Todd. *Psychological Review*, *71*(6), 528–530. https://doi.org/10.1037/h0047061

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved Rˆ for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, *16*(2). https://doi.org/10.1214/20-BA1221

Ver Hoef, J. M. (2012). Who invented the delta method? *American Statistician*, *66*(2), 124–127. https://doi.org/10.1080/00031305.2012.687494

Vrij, A. (2008). Nonverbal dominance versus verbal accuracy in lie detection: A plea to change police practice. *Criminal Justice and Behavior*, *35*(10), 1323–1336. https://doi.org/10.1177/0093854808321530

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). *MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers* (arXiv:2002.10957). arXiv. https://doi.org/10.48550/ARXIV.2002.10957

Wenzler, S., Levine, S., Van Dick, R., Oertel-Knöchel, V., & Aviezer, H. (2016). Beyond pleasure and pain: Facial expression ambiguity in adults and children during intense situations. *Emotion*, *16*(6), 807–814. https://doi.org/10.1037/emo0000185

Wilson, A., & Adams, R. (2013). Gaussian process kernels for pattern discovery and extrapolation. *Proceedings of the 30th International Conference on Machine Learning*, 1067–1075. https://proceedings.mlr.press/v28/wilson13.html

Winfrey, O. (Director). (2021, March 7). *Oprah with Harry and Meghan* [Broadcast]. Harpo Productions.

Żelasko, P., Pappagari, R., & Dehak, N. (2021). What helps transformers recognize conversational structure? Importance of context, punctuation, and labels in dialog act recognition. *Transactions of the Association for Computational Linguistics*, *9*, 1163–1179. https://doi.org/10.1162/tacl_a_00420

Zhou, K., Ethayarajh, K., Card, D., & Jurafsky, D. (2022). Problems with cosine as a measure of embedding similarity for high frequency words. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 401–423. https://doi.org/10.18653/v1/2022.acl-short.45