UNIVERSITY OF CALIFORNIA

Los Angeles

Variation and Uniformity in Mental State Talk Across Three Languages

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Anthropology

by

Andrew Marcus Smith

2024

ABSTRACT OF THE DISSERTATION

Variation and Uniformity in Mental State Talk Across Three Languages

by

Andrew Marcus Smith

Doctor of Philosophy in Anthropology

University of California, Los Angeles, 2024

Professor H. Clark Barrett, Co-Chair

Professor Erica A. Cartmill, Co-Chair

Does the way we talk about other people's minds depend on the language we speak? This dissertation explores this question by developing and applying a novel methodology to systematically collect and analyze standardized corpora of speech samples about others' minds. Using this approach, I created a cross-linguistic corpus from English speakers in the United States, Mandarin speakers in China, and Arabic speakers in Morocco. This corpus was used across three studies to determine whether the frequency of mental state talk varied across languages and whether individual variation in the frequency of mental state talk was related to an underlying dimension of social cognition known as mindreading—the ability to infer others' mental states. The first study analyzed the production of eight key mental state verbs theorized to be critical for mindreading development across field sites. However, this narrow focus overlooked much of the mental state lexicon. The second study addressed this limitation by coding all mental state terms in the corpus as identified by native speakers of each language. The third study examined whether participants' frequency of mental state talk predicted their

performance on the Reading the Mind in the Eyes Test (RMET), a widely used measure of mindreading ability, and whether this relationship differed across languages. Three key findings emerged. First, the frequency of mental state talk was largely consistent across cultural-linguistic contexts, suggesting it may occur at a relatively fixed rate that is independent of cultural and linguistic variation. Second, mental state talk frequency significantly predicted RMET performance, though participant talkativeness was a slightly stronger predictor. Third, both factors were consistent positive predictors of RMET scores across all field sites. These findings suggest that the relationship between mental state talk and mindreading competence is less influenced by cross-cultural or cross-linguistic differences than previously thought. They also emphasize the importance of considering not only the specific content of mental state talk but also the broader linguistic context when studying social cognition. This work advances a more nuanced understanding of the interplay between language, culture, and our ability to understand others' minds.

The dissertation of Andrew Marcus Smith is approved.

Gregory A. Bryant

Richard Alan Clarke Dale

Daniel Fessler

Erica A. Cartmill, Committee Co-Chair

H. Clark Barrett, Committee Co-Chair

University of California, Los Angeles

2024

This dissertation is for my family and friends, the city of Los Angeles, Nick's Coffee Shop and Ballona Creek; it's for Zach and Maya and Syd and Cass; it's for my mom and my dad and my dad; it's for Theo and Kotrina and Renee; for Val and Sam and Aviv and EJ and Raina; it's for the Piaggio Scooter I never got licensed to drive; it's for the Piaggio Scooter I drove nonetheless; it's for the hubris I felt thinking the only reason people in LA couldn't drive in the rain was because it happened so infrequently and it's for the same Piaggio Scooter I was riding when I sprained my ankle taking a sharp turn… in the rain; it's for living in K-town; it's for Louie and Sarge; it's for telling Kat and Derek to pound sand; it's for the virtue of forgiveness and the cultivation of compassion; it's for Lee and Eugene and Steven and Nadav; it's for finally, FINALLY moving out of K-town; it's for learning to surf; it's for your brother driving six hours from Phoenix to LA every month and not vice versa because he owns a car and you don't, even though you *could* take the bus to Arizona; it's for Shabbat dinner with your camp friends; it's for brewing beer and making kimchi; it's for failure and success and for relearning what those words actually mean; it's for getting your heart broken and it's for thinking your best days are behind you; it's for the impossible ember of hope that saw you through your darkest night and it's for the sweet orange cat who curled himself around your heart to protect it when you yourself could not; it's for coming home to Ithaca, a decade closed and shut, with tales of Polyphemus and the War on your tongue; but most of all, it's for falling in love. It's for Abby. It's for finding your besherte in the foothills of the Angeles National Forest. It's for laying your soul bare before someone who sees you, completely. It's for the sound of her laughter. It's for you and me, you and me, nobody, baby, but you and me. It's for thrifting and rockhounding and the Island Fox. It's for realizing, finally, that your best days are only just beginning. In short, it's for you – all of you, without whom it would have been impossible.

"What is to give light must endure burning." – Viktor Frankl

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

ACKNOWLEDGEMENTS

The author thanks the John Templeton Foundation for providing the funding to support the Geography of Philosophy Project, of which the research in this dissertation constitutes a part. Abdellatif Bencherifa and Salma Tber contributed to this work through their efforts recruiting and running participants in Morocco. Xiaofei Liu contributed similarly in China. The author extends his significant gratitude to his team of research assistants who helped correct transcripts of participant speech samples and code them for subsequent data analysis. This work was completed by Ali Ashkanani and Khaoula Assabar for the Moroccan dataset and Yuexin Wren Xu and Ziyi Meng for the Chinese dataset. A special thanks to the team of research assistants who recruited and ran participants at UCLA in addition to correcting and coding transcripts. Ahmet Dikyurt, Thu Phan, Leyi Qiu, Tegan Roberts, and Matthew Tran contributed to the current work in this way. Special thanks to Stephen Stich, Edouard Machery, and H. Clark Barrett, the Principal Investigators on the Geography of Philosophy Project, for making this work possible by including the studies in this dissertation as part of their broad, international research effort. The author would like to extend his sincere gratitude to Gregory A. Bryant, Richard Alan Clarke Dale, and Daniel M.T. Fessler for their support and incisive feedback, especially in writing the first chapter of this work. Thanks also to the UCLA Center for Behavior, Evolution, and Culture and its members, without whose friendship and community this work could not have happened. And finally, the author extends his warmest, sincerest, and most grateful thanks to Erica A. Cartmill and H. Clark Barrett, whose mentorship, support, and brilliance were a precious gift to the author’s intellectual development. Throughout his time as a graduate student, they have incubated his curiosity and pushed him to think bigger, and for that he well be forever thankful.

VITA / BIOGRAPHICAL SKETCH

Andrew Marcus Smith was awarded a Bachelor of Arts in Cognitive Science from Rutgers University, graduating summa cum laude. While at Rutgers he was inducted as a member into Phi Beta Kappa and was honored as a Paul Robeson Centennial Scholar for completion of his senior thesis. He was further awarded a Master of Arts in Psychology from UCLA. He has worked in a wide range of research contexts, with scholarly contributions in the fields of neuroscience, anthropology, cross-cultural psychology, and medical practice. Papers on which he has been featured as an author have been published in Nature Human Behavior, Proceedings of the Royal Society B, the Journal of Experimental Psychology: Animal Behavior Processes, Behavioural brain research, the Neurobiology of learning and memory, and Cureus. Andrew has received several fellowships and awards to support and recognize his research efforts, including the Blum Center Summer Scholarship, the Foreign Language and Area Studies Fellowship, and the Graduate Summer Research Mentorship award. In his time as a graduate student at UCLA, Andrew has held a number of Graduate Student Researcher Positions, including as the Haines Hall Digital Media Lab Research Assistant for two years and as a Research Manager for several projects undertaken by the Leadership Education in Neurodevelopmental and Related Disabilities program, known otherwise as UC-LEND. He has also held a variety of Academic Student Employee positions, serving as a teaching assistant for courses in the Departments of Psychology, Anthropology, Communication, and Molecular, Cell, and Developmental Biology.

Chapter 1: The Empirical and Theoretical State of the Art in the Study of Mindreading

Introduction

This dissertation seeks to determine whether there exists cross-linguistic variation in the frequency of mental state talk and whether such variation, if found, covaries with the mindreading capacity. To this end, a novel and generalizable methodology was developed for the production of systematic, standardized corpora of speech samples about the minds of third parties through elicited narrative retellings of custom-made video stimuli depicting naturalistic, everyday social interactions. This methodology was used to generate a cross-linguistic corpus of American English, Moroccan Arabic, and Mandarin Chinese speech which was then coded according to two distinct schemes. The first was based on a bank of 8 mental state verbs theorized by some scholars to bear a privileged functional relationship with mindreading, over and above that of other mentalistic verbs, adjectives, and nouns (Gleitman, 1990; Shatz et al., 1983). This coding scheme was designed to capture all instances where these verbs referred to the mental states of third parties. Counts of these instances were then used to assess whether speakers of English, Arabic, and Chinese differed in the frequency of their talk about the mental states of third parties. The second coding scheme was based on theoretically-driven skepticism about the privileged status of these 8 mental state verbs and aimed to capture any and all mentalistic words of any grammatical category. This was achieved by training fluent, native speakers of each language on an operational definition of mental states and using their linguistic insight to categorize each and every word in the corpus as a mental state or not. Individual instances, or tokens, of candidate mental state words were checked in context to ensure they referred to the mental states of third parties. Token counts were then used to assess whether the relative cross-linguistic frequencies of third-party mental state talk replicated those observed when using the first coding scheme. Finally, per-participant counts of mental state word tokens were calculated according to each of the two coding schemes. These values

were then coupled with participant performance on the Reading the Mind in the Eyes Test, or RMET (Baron-Cohen et al., 2001), a widely used measure of mindreading ability in adults. Each set of per-participant mental state word counts was then modeled as a function of participant RMET scores to determine whether there existed a relationship between the frequency of participants' talk about the minds of others and their mindreading ability. Additionally, separate models for counts produced using each of the two coding schemes allowed it to be determined whether the strength of this relationship varied when focusing on just the 8 theoretically important mental state verbs as opposed to all mental state terms. To these ends, this chapter reviews the extant literature on mindreading, language, communication, and their intersections to map the empirical landscape and motivate the studies comprising this dissertation and the questions they will help to answer. This chapter argues that these questions are both "low-hanging fruit" within the problem space and that their answers are fundamental to resolving broader questions about the relationship between social cognition and language. In mapping out this literature, I underscore the need for a methodology of the sort developed in this dissertation and highlight the urgency of its application to the targeted empirical problem this dissertation addresses – namely, whether mental state talk exhibits cross-linguistic variation and if it covaries with the mindreading capacity.

Mindreading

Mindreading refers to the ability to impute the mental states of other agents, including their perceptions, emotions, intentions, and attitudes. The function of such a cognitive capacity may seem relatively straightforward and its scope circumscribed, but such conclusions underestimate the complexity of the problems it solves and the breadth of information it draws upon to generate such solutions. Consider, for example, the following two scenarios. In the first, you witness an unfamiliar individual standing on a packed subway car, their arm raised above their head to grasp the handrail and their face mere inches from the wall of the car's interior. In

the second, you witness the same unfamiliar individual holding their body in the exact same position, their arm raised above their head and their face mere inches from the wall. Now, however, the two of you are not located on a busy subway car but an empty subway platform. For many readers, imagining oneself in the latter of these two scenarios may cause some unease or discomfort, while the former may be so utterly banal as to elicit no emotional response whatsoever. In the latter, one might imagine themselves getting up from their seat and walking toward the far end of the platform to avoid this unknown individual, perhaps looking for a nearby exit while removing an earbud to better monitor their surroundings. In the former, one might imagine themselves briefly noticing this unknown individual on the car before turning their attention back to the book in which they'd been absorbed.

The question, then, is why do these two scenarios elicit such different thoughts, feelings, and behaviors despite their broad commonalities? Trivially, it is because of the features that differ across them. More meaningfully, it is the impact those differences have on the mental states imputed to the unknown individual by the mindreading capacity. Across these two cases, the mindreading system draws upon perceptual inputs to represent the context and agents in one's immediate environs (A. Clark, 2013; Gilbert et al., 2015; Leslie, 1994). It also draws upon long-term and short-term memory to serve up learned associations between contexts and typical agentive behavior in those contexts, as well as learned representations of the mental states likely held by and motivating the behavior of agents in those contexts (A. Clark, 2013; Emery & Clayton, 2001; W. S. Hall et al., 1981; McCabe et al., 2000; Parrigon et al., 2017; Tomasello & Carpenter, 2007). Biases for interpreting the behavior of novel agents, like the intentional stance, or the tendency to assume the existence of a motive behind behavior that might otherwise be opaque, also serve up information to working memory to impute mental states in the current context (Gergely et al., 1995; Southgate et al., 2007). All of this information is used by the mindreading capacity to impute the mental states of the people around oneself

and thereby make predictions about their behavior. To the extent their behavior is relevant to one's own interests, the mindreading capacity allows us to predicate our own actions on the anticipated moves of those around us.

Critically, the outputs of the mindreading system are probabilistic in nature and depend on the quality of the information fed in. Where the mental states imputed to another person are more certain, so too are their anticipated behaviors and so too is the certainty with which one's subsequent behavior will serve their own interests. Where they are less certain, that uncertainty feeds into predictions further down the causal chain. If another person's mental state is uncertain, we are less sure of how they will behave and less sure of how their behavior will impact ourselves. In the first of the two scenarios, we know why the individual's arm is up. We know why their face is so close to the wall of the subway car – there is no room and so they have likely taken whatever space is available to them. They are likely holding the handrail to stabilize themselves, as there were no seats available and they had to stand. They might not necessarily want to have their arm up, or to have their face so close to the wall of the subway car, but these represent contextual constraints on their likely goal, which is to commute. They will most probably mind their own business and interact minimally with the other commuters on the subway car. This will likely have very little impact on our own interests, and as such we can more or less confidently cease paying attention to them. The same cannot be said of the second scenario. There are no contextual constraints that concisely explain the unknown individual's stance, no obvious reasons to be stand with their face just a few inches from the wall with their arm elevated. Without a clear rationale for their current behavior, it is difficult to anticipate what they will do next and whether it will impact our own interests. This uncertainty presents a risk, and we can thus contingently shape our behavior to mitigate it.

The importance of this capacity should be clear – the mindreading ability fundamentally undergirds the way in which we navigate the world because the world is one defined by social

interaction. Other agents, human, animal, or otherwise, can help us or they can harm us. It is thus not an overstatement to say that the ability to reliably anticipate what someone thinks, which course of action they'll take, and whether it will hurt or harm ourselves is central to our survival. Despite the richness of this view, it warrants noting that it is one that contrasts with other, narrower, and more broadly embraced perspectives on the mindreading capacity, though it has been criticized (P. Bloom & German, 2000). Many scholars, for example, have treated the mindreading capacity as identical to or interchangeable with just the representation of others' false beliefs. In the context of the mindreading capacity, false beliefs are those held by an individual which do not accurately represent the state of affairs to which they correspond. For example, I may grab the tube on my kitchen sink believing it contains toothpaste, only to discover that it is in fact ointment upon brushing my teeth. In this case, I held the false belief that the ointment was toothpaste. As another example, I might place my jacket on a chair upon entering my home before heading to my bedroom. While in the bedroom, my partner might hang my jacket in the closet. Upon exiting the bedroom, I am surprised to see that my jacket is not where I left it. In this case, the false belief I held was that my jacket was still resting on the chair where I had left it.

The equivocation of mindreading with the representation of false beliefs is likely a consequence of the history of mindreading research, wherein false beliefs were first identified as providing a mechanism through which to probe individuals' second-order representations, or representations of another agent's representations (P. Bloom & German, 2000; Wimmer & Perner, 1983). Specifically, a number of tasks which required participants to track the false beliefs of another agent were developed by researchers in the early 1980s to explore the age at which this ability first emerged in children (Baron-Cohen et al., 1985; Gopnik & Astington, 1988; Wimmer & Perner, 1983). These tests, known broadly as False Belief Tests, take a number of forms and allowed researchers to examine a variety of false beliefs. The paradigmatic version of

the False Belief Test, however, presented to participants a scene with two agents, a box, and a basket. The first agent was shown playing with an object, whereupon the agent placed the object in one of the hiding locations. The second agent was present during the hiding event. After the first agent had placed the object in one of the locations, they left the scene. The second agent then relocated the object from the first location to the second in the first agent's absence. The second agent then left the scene. The first agent then returned. At this point, the participant was asked three questions assessing their understanding of the previous state of affairs (object in location 1), their understanding of the current state of affairs (object in location 2), and their understanding of the first agent's belief about the location of the object (object in location 1). Children under the age of four were found to reliably answer questions about the current and previous state of affairs accurately but were not found to reliably answer the false belief question accurately. Under this approach, children were found not to reliably answer the false-belief question correctly until after the age of four.

While the ability to represent others' false beliefs (i.e., to represent a representation which differs from reality and is held by a second- or third-party) unequivocally constitutes a component of mindreading, it is far from the totality of sociocognitive functions subsumed by mindreading. Though much of the early empirical work on mindreading focused on false belief, subsequent theoretical and empirical developments have moved away from this narrow understanding. Current empirical and theoretical understanding makes clear that false belief representation is just one of the many sociocognitive functions subsumed by the mindreading capacity. Nevertheless, it is from this representation-focused foundation in cognitive science that research on mindreading began in earnest.

Mindreading across human populations

Infant research

In the early 2000s, two studies demonstrated that children as young as 15-months old could pass a modified false belief task which did not impose the same linguistic and pragmatic demands of the measure as originally designed (Onishi & Baillargeon, 2005). In effect, Onishi and Baillargeon showed that infants could “spontaneously” attribute false beliefs, or do so without needing to answer any questions. This was in contrast to the test’s earlier “elicited” forms which required children to answer several questions. Moreover, it was found that children as young as twenty-five months old could form expectations about an agent's future actions based on its earlier goal-directed behavior and its present perceptual access to the goal (Southgate, Senju, & Csibra, 2007). These studies showed that when an agent who ought to have a particular belief or desire (i.e., a false belief or a desire concordant with earlier behavior) behaved in a way that violated an infant's expectations, they looked at the study stimulus for a longer period of time than when the agent behaved according to their expectations. To the extent that looking time indexes interest, and to the extent that unexpected events are more interesting, these data suggest that infants may have an implicit or automatic capacity to track false beliefs which does not rely upon linguistically- or pragmatically-mediated reasoning abilities. Infants were increasingly shown to have a complex understanding of others' intentions and motivations that extended beyond just the ability to track others’ false beliefs. Gergely et al. (1995) found that infants as young as 12 months old could identify an agent's goal and interpret its actions in terms of that goal, attributing intentional mental states like beliefs and desires to others (Gergely et al., 1995). It is perhaps unsurprising that infants as young as six months old were subsequently shown to engage in a number of behaviors that undergird the attribution of mental states to others, including the ability to detect and follow others’ gaze, (D’Entremont et

al., 1997; Farroni et al., 2002; Tomasello et al., 2007), to perceive biological motion (Simion et al., 2008), to produce gestures directing the attention of others (Cochet et al., 2017; Liebal et al., 2009), and to automatically encode others' beliefs (Kovács et al., 2010). Beyond merely imputing the mental states of conspecifics at an early age, it has been found that children as young as 18 months old can use these capabilities to direct cooperative and collaborative efforts, as evidenced by their early competence in tasks of joint attention (Carpenter & Tomasello, 1995). The capacity for joint attention involves the intentional sharing of attention with another individual, a capacity that requires an understanding of the fact that seeing leads to knowing, and that by directing someone's attention to an object both agents are seeing and knowing the same thing, together. This is a massively complex sociocognitive behavior which nevertheless emerges early in human development.

Cross-cultural research

Research in small-scale societies has contributed greatly to our understanding of mindreading and its development within and across social ecologies. Baka children in a Cameroonian community were found to pass the False Belief task reliably, albeit at a slightly later age (between the ages of 4 and 5) than children in American and European contexts (between the ages of 3 and 4). (Avis & Harris, 1991). Two more recent cross-cultural studies employing a series of both spontaneous and elicited false-belief tasks, respectively, replicated this finding, pointing toward a universal competence in representing the beliefs and desires of others (Barrett et al., 2013; Callaghan et al., 2005; Slaughter & Perez-Zapata, 2014). Work in urban and rural communities in Vanuatu has suggested considerable differences in the age at which competence in the False Belief task is reached, suggesting important social and cultural influences on the development of mindreading (Dixson et al., 2017). A meta-analysis of false-belief understanding across communities in mainland China, Hong Kong, the United States, and Canada has shown that while there exist parallels in the development of mindreading, there are

significant differences in the timing such that the age at which children achieve competence varies by as much as two years (Liu et al., 2008). Beyond these cross-cultural differences in early mindreading competence, cross-cultural differences in practices related to mindreading have been documented among adults in a number of societies in the South Pacific, including Samoa, Papua New Guinea, and Vanuatu (Dixson et al., 2017; Robbins & Rumsey, 2008; Schieffelin, 2008). Linguistic and cognitive anthropologists have posited that in these cultures, there are norms which prohibit the attribution of mental states to others, the existence of which may meaningfully reduce the frequency with which such attributions occur (Robbins & Rumsey, 2008). Subsequent research has shown that Ni-Vanuatu children up to the age of 14 years old do not perform above chance on the False Belief Test, depending on whether they were recruited from a rural or urban context (Dixson et al., 2017). Taken together, these data could constitute evidence of a role for enculturation into regimes of interaction and social practice in shaping the mindreading capacity. Beyond differences between cultural contexts, it has been shown that adult participants tend to respond more rapidly and more accurately when making mental state attributions to individuals within their own cultures as opposed to across cultural contexts (Perez-Zapata et al., 2016). While these findings have proven crucial to our understanding of the ways in which mindreading varies across human populations, fruitful discussions on mindreading have taken these findings to task and have generated data increasing the breadth of species to which some mindreading phenomena can be attributed. These data delineate those components of mindreading argued to be unique to human beings and are crucial to the development of theories permitting interactions between language and mindreading.

Mindreading across species

Having detailed some of the extant data on human mindreading, I now review the comparative cognition literature to inventory both those elements of the mindreading capacity

that exist widely across the animal kingdom and those that are derived in the human lineage, clarifying the components shared with other species and aiding in the identification of elements that interact bidirectionally with language. Though ancient and phylogenetically ubiquitous cognitive abilities, like vision and attention, almost certainly interact with and shape at least some aspects of language, there is less evidence to support the claim that language restructures or shapes aspects of more highly conserved elements of cognition (Firestone & Scholl, 2016). While it is an assumption that language may interact with and shape uniquely human components of mindreading more strongly than widely distributed and highly-phylogenetically conserved ones, research has shown that phenotypic plasticity can be selected for under conditions of variability in the pressures relevant to a given trait (Gilbert et al., 2015; Levis & Pfennig, 2016). To the extent that uniquely human mindreading serves to navigate interactions with other agents, and to the extent that such agents exist in highly variable cultural contexts which differentially condition their behavior, while using highly variable linguistic systems to mediate their interactions, it stands to reason that uniquely human mindreading may exhibit adaptive plasticity in response to these variables.

Non-human primates

Extensive work has been done to examine non-human primate mindreading abilities which can shed light on the uniqueness of human mindreading. To date, several components of mindreading have been documented among non-human primates. Most recently, bonobos, chimpanzees, and orangutans were shown to track false beliefs in a paradigm analogous to looking-time studies used with infants (Kano et al., 2017; Krupenye et al., 2016). Chimpanzees and bonobos have also been shown to engage in gaze following (Tomasello et al., 2007), to understand that sight plays a role in establishing knowledge (Hare et al., 2000), to use such knowledge in the service of deceiving conspecifics (Whiten & Byrne, 1997), and to differentiate between accidental and intentional action (Call & Tomasello, 1998). Nevertheless, there are a

variety of domains related to mindreading with which non-human primates struggle. These domains are not necessarily mindreading *qua* mindreading, but their success may be predicated on components of mindreading absent in non-human primates. Most great apes tend to struggle in cooperative tasks that require shared intentionality (Carpenter & Tomasello, 1995; Moll & Tomasello, 2007). Under its standard conception, shared intentionality, "…is a theoretical construct that refers to a suite of abilities that enable coordinated, collaborative interactions, and claims that the mechanism to obtain these skills reside in the sharing of mental states, such as attention and goals…" (Persson et al., 2023; Tomasello, 2019). Some of these specific tasks include joint attention as opposed to simple gaze following, cooperation as opposed to social manipulation, collaboration as opposed to mere group activity, and deliberate teaching as opposed to simple social learning (Tomasello & Carpenter, 2007). Additionally, they do not appear to engage in triadic joint attention (Tomonaga et al., 2004) or gaze-checking (though see Bräuer et al., 2005 for some evidence to suggest otherwise), both of which are early-developing aspects of human mindreading (Carpenter & Tomasello, 1995). Chimpanzees, unlike humans, do not copy mechanically and causally ineffectual behaviors that have been demonstrated to them. This may reflect a human bias to attribute intentionality to both the *way* and the *why* of an action (Lyons et al., 2007). In the absence of these skills, non-human primates appear not to be capable of the kinds of behaviors that facilitate more explicit, and potentially uniquely human, aspects of mindreading.

Corvids and caching behavior

Comparative questions about mindreading have been extended to animals other than our nearest living relatives. While data from non-human primates may suggest something about the features of human mindreading rooted in homologous mechanisms shared with other species, these same data do not definitively answer questions about the kinds of extrinsic, ecological factors that might have pushed ancestral primates in this direction in the first place.

To that end, a great deal of research has focused on the mindreading capacities of corvids in an effort to assess both the ecological and social factors that may select for mindreading. Many corvids are caching species, which means their feeding behavior involves locating, storing, and relocating food items for later. Because caches are left unguarded, they are vulnerable to raiding. This ostensibly presents selection pressures for birds to monitor the caching behavior of conspecifics, and in turn, to track whether conspecifics are monitoring their own caching.

A number of compelling studies have demonstrated that corvids who cached food items in the presence of conspecifics were significantly more likely to re-cache them when given the opportunity to recover the item in private later on than corvids who cached food items in the absence of conspecifics. (Emery & Clayton, 2001). Some researchers argued that these findings did not unequivocally demonstrate important aspects of mindreading and that corvids may merely be employing behavioral rules or heuristics like "re-cache your food if a competitor is present initially". (Butterfill & Apperly, 2013). Later studies provided more definitive evidence of mindreading among corvids by controlling for associative cues that may trigger such behavioral rules or heuristics (Bugnyar et al., 2016). However, whether this distinction meaningfully carves out mindreading from something else is a position worth treating with some skepticism. To the extent that *all* mindreading extrapolates from the regularity of certain perceptual cues, the difference between "behavioral rules and heuristics" and mindreading might be one of quantity rather than quality, without a clear point at which one can be said to switch over to the other and vice versa (Barrett, 2015; Whiten, 1996).

Carnivores

While data generated across taxa have highlighted some of the ecological contexts in which mindreading may have undergone positive selection, carnivory presents another avenue through which it may have evolved (Barrett, 2005). Many of the competencies involved in mindreading bear on hunting dynamics. Prey animals are well-served by detecting the gaze of

potential predators and tracking if it is following them, as these constitute potential cues to predation. Additionally, predators and prey alike need to represent agents as distinct from other objects in the world to ensure their survival. Though constrained in scope, representations of this nature are critical to the success of both predators and prey.

Despite the centrality of mindreading to predator-prey interactions, carnivores have been relatively underrepresented in studies of animal social cognition to date (Benson-Amram et al., 2023). This underrepresentation applies to carnivores understood both as species within the order *Carnivora* as well as non-*Carnivoran* species that consume a primarily or exclusively carnivorous diet. Nevertheless, research on true *Carnivorans* like dogs (Huber & Lonardo, 2023), wolves (Range & Virányi, 2011; Virányi et al., 2008), hyenas (Holekamp, 2007), and cats (Quaranta et al., 2020), as well as research on carnivorous non-*Carnivorans* like cetaceans (Davies & Garcia-Pelegrin, 2023) and reptiles (Doody et al., 2013) have both shed light on the role played by carnivory in shaping the mindreading capacity. The preponderance of evidence suggests that while carnivory *qua* carnivory may select for more basal mindreading competencies like gaze and agency detection, more complex mindreading phenomena like joint attention tend to emerge in the context of particular kinds of social organization and particular patterns of social interaction (Udell et al., 2011).

Where there are regular, structured, and stable social interactions with conspecifics and where an individual's fitness is related to their ability to navigate their social world, pressures are introduced to better read the minds of one's interlocutors. Wolves, hyenas, and cetaceans all exhibit societies structured in this way (Davies & Garcia-Pelegrin, 2023; Holekamp, 2007; Range & Virányi, 2011; Virányi et al., 2008). However, rich sociocognitive abilities may also emerge in response to selection pressures other than sociality outright. For example, regular interaction with human beings may result in the enrichment of these abilities among species that are less gregarious, like domestic cats (Quaranta et al., 2020).

Fitting the evidence together

Collectively, these data highlight a number of mindreading homologs and analogs of across phyla. The ability to detect and follow gaze has been observed across the broadest range of species, including corvids, non-human primates, carnivores, ungulates, and even red-footed tortoises (Wilkinson et al., 2010). The species that has demonstrated the most human-like mindreading abilities is arguably chimpanzees, which appear to relate the gaze of others to their states of knowledge (Hare et al., 2000), detect biological motion, and differentiate between accidental and intentional action (Call & Tomasello, 1998).

Despite a rich shared capacity for mindreading across phyla, human beings are unique in the depth and breadth of our mindreading, mediated in part and elaborated by its relationship with both language and culture (De Rosnay et al., 2014; Heyes, 2018; Lagattuta et al., 2010; K. Milligan et al., 2007). Human beings track recursive belief structures with some ease and do not automatically assume transitivity across recursive layers. John may believe that snow is green, and I may believe that John believes that snow is green, but I do not necessarily believe that snow is green. These kinds of propositional attitudes are perhaps unique to human beings, and the predictions mindreading of this kind affords may be distinctly human. If I know that Marion has disliked every horror movie we've ever seen together, I can use this knowledge of her preferences to predict the outcome of a given course of action and to plan my behavior contingently upon this mental model. If I were to download a horror movie for our next movie night, she would be angry because either I know she doesn't like them, and I don't care or I forgot and have failed to keep track of her preferences. As such, I ought not download another horror movie for when I see her next. The extent to which we can build such mental models of the attitudes, emotions, intentions, and perceptions of others varies as a function of our experience with them.

The capacity to build models of others' minds may mark the beginning of uniquely human mindreading rather than its end, the outputs of such representational models feeding forward into other uniquely human forms of cognition. For example, ostensive signals, or communicative signals which indicate a communicator's intention to share information, can be targeted to specific individuals by conditioning their production on representational models of their knowledge (Scott-Phillips, 2014; Sperber & Wilson, 2001). Such encrypted signals could not be produced unless these mental state representations played a role in the production of communicative signals that optimize Gricean communicative maxims, a set of rules followed by people in order for communication to occur cooperatively between individuals and for utterances to be understood. These rules are to be informative, to be truthful, to be relevant, and to be clear. Respectively, these constitute the maxims of quantity, quality, relation, and manner (Misyak et al., 2016; Okanda et al., 2015). Where an apparent violation of these maxims occurs, as in the case of an encrypted signal (which may appear to violate the maxim of manner or relation), the receiver may infer that the signal was designed to be interpretable only to themselves. Otherwise, the signal would have been clearer or more relevant if the signaler was indeed optimizing Gricean communicative maxims.

Moreover, the outputs of representational models of others' minds might feed into mechanisms of moral decision-making in uniquely human ways. When an interlocutor's behavior imposes some negative cost, one's representation of the interlocutor's mental state may serve to calculate the relative probabilities that the cost constituted an error as opposed to malice. Moral decision-makers can then act on these outputs. Though such representations may seem intrinsically general and thus richly flexible, it is not straightforwardly the case that generality grants flexibility for free. Rather, these mechanisms exist in concert with the rest of the cognitive apparatus. The connection of mental state representations to other cognitive capacities, like language, constitute critical factors to consider in the evolution of human

mindreading and have likely shaped this apparent generality. With this argument in mind, and with my review of the empirical data complete, I now turn toward some of the theoretical perspectives on mindreading that have emerged in the past several decades and evaluate them in terms of how readily they accommodate interactions between mindreading and language.

Mindreading across theoretical perspectives

Traditionally, theories of mindreading have been characterized as belonging to one or the other of two major theoretical positions – theory-theories or simulation-theories. Theory-theories posit that a change occurs in the conceptual structure, or theory, children use to understand and explain the behavior of others over the course of early childhood such that their predictions become more accurate over time (Gopnik & Astington, 1988; Gopnik & Wellman, 1992; Wimmer & Perner, 1983). Simulation-theories argue instead that the way children come to understand and explain the behavior of others is by imagining themselves in a given circumstance and attributing to others the mental states they experience (Gallese & Goldman, 1998). More contemporary theories include those of Apperly and Butterfill (2009), Heyes (2018), Leslie (1994), Baron-Cohen (1997a), and Nichols and Stich (2003). Each of these theories represent departures from theory-theory and simulation-theory accounts and introduce additional mechanisms to explain how the mindreading capacity operates. In an appeal to arguments of the type made by Tversky and Kahneman (1974), Apperly and Butterfill (2009) suggest that mindreading problems are served by one or more Type 1 mechanisms that rapidly and inflexibly produce low-cost outputs. These mechanisms are fast, intuitive, and largely automatic, processing information quickly and often without conscious awareness or control. Due to their speed and automaticity, the outputs of Type 1 mechanisms can be prone to bias and errors. That there may be one or more Type 1 mechanisms is a notion consistent with massive cognitive modularity (Fodor, 1983).

Apperly and Butterfil (2009) also posit that there also exists a Type 2 system in the domain of mindreading, where effortful, costly, and slower cognitive processes can operate on and evaluate the accuracy of Type 1 outputs. Type 2 processes are often slower, more deliberate, and require conscious effort for complex reasoning and decision-making. These mechanisms require more cognitive effort, are under conscious control by individuals, and are generally more accurate when used to carefully evaluate information. Heyes's (2018) theory marks a significant departure from other theories and argues that the mindreading capacity is best conceived of as a "cognitive gadget," a term chosen to underscore that mindreading is a learned and socially constructed tool that may vary across human populations. Under Heyes's theory, mindreading is not innate and it is learned across development through culture.

In contrast, Leslie (1994) struck a strong claim to the structure and operation of the mindreading system in an attempt to explain a phenomenon he called "Agency". Under Leslie's account, "Agency" was defined as a conceptual primitive composed of three distinct domains of knowledge, each of which tracked distinct properties of the world and each of which was processed by a corresponding cognitive subsystem. These domains of knowledge are mechanical "Agency", actional "Agency", and attitudinal "Agency", which correspond respectively to the mechanical properties of agents, the goal-directedness of the actions produced by agents, and the mental states motivating those actions. Leslie posited that the conceptual primitive of "Agency", composed of these three parts, emerges from the interplay of at least two modules, a Theory of Body mechanism (ToBy) and a Theory of Mind Mechanism (ToMM). Specifically, these two modules supported domain-specific learning which served as the foundation upon which the concept of "Agency" was built. Leslie thus staked a claim to the structure of the mindreading system in order to explain these levels of "Agency". Unlike Leslie, Baron-Cohen's major functional presumption is that human beings need to interpret and predict action (1997a). Human beings need to engage in both dyadic and triadic interactions to facilitate

interaction and direct joint attention on shared targets of interest. As such, organisms need to be able to identify the volitional states, gaze, and mental states of their interlocutors, as well as to ascertain whether they are both attending to the same stimuli. Baron-Cohen suggested a system of mindreading based on four modules – an intentionality detector (ID), an eye direction detector (EDD), a shared attention mechanism (SAM), and a Theory of Mind Mechanism.

Last among the theories to review is that of Nichols and Stich (2003) who sought to account for features of the available empirical data on mindreading which were at the time poorly explained. These data primarily concern the ability of children to engage in pretend play, a behavior that involves complex representational skills. The authors take as a conceit that there are two broad kinds of mental state representations that structure decision-making and influence behavior – beliefs and desires. Nichols and Stitch also propose the existence of the Belief Box, the Desire Box, and the Possible Worlds Box (PWB). Both the PWB and the Belief Box receive input from an Inference Mechanism that serves to derive conclusions from an existing set of beliefs. An additional mechanism in the model is the UpDater, which provides new beliefs and premises to the inference mechanism and thus allows for feedback and elaboration of the set of beliefs currently held in the Belief Box. Such feedback loops allow the system to integrate new information that bears on the current circumstances. The Script Elaborator is another mechanism posited that allows specification of free parameters in the pretend premises that are not themselves logically constrained, which may support inferential processes about the mental states of one's interlocutors which are not inconsistent with one's current understanding of the content of their minds. Thus, Nichols and Stitch suggest that belief and desire representations, coupled with these novel mechanisms, account sufficiently for as-of-yet unexplained features of pretend play, including elaboration of pretence and navigation around the logical constraints imposed by prior elaborations. Having reviewed these theoretical positions, I now synthesize their findings to identify gaps and offer a view lending credence to

perspectives that take seriously the connection between mindreading and language. Namely, I present a view of mindreading that allows for cross-cultural and cross-linguistic variation in its manifestation while still allowing for there to exist shared, universal components.

Synthesizing data and perspectives with an eye towards language

Taken together, these data suggest several points to consider. First, mindreading appears not to be a singular competence, but the emergent outcome of a varied suite of mechanisms and abilities. Indeed, several methodological papers have emphasized the importance of developing tools for assessing the dissociable parts of mindreading to better examine its subcomponents, an endeavor that may be enriched by contributions from the comparative cognition literature as well (P. Bloom & German, 2000; Turner & Felisberti, 2017; Wellman & Liu, 2004; White et al., 2009). Thus, it may be inappropriate to commit to theories that seek to minimize mindreading's moving parts. Second, false belief is likely not the gold-standard metric by which mindreading ought to be measured. As such, the research framework from which it emerged should be weighed in proportion (P. Bloom & German, 2000). Indeed, adults on the autism spectrum pass the false belief task with ease (Castelli et al., 2002), despite this cluster of neurological differences having been characterized as a deficit in mindreading. If false-belief reasoning is the core component of mindreading, then this finding begs explanation.

Third, it appears to be the case that some competencies emerge in early infancy at very nearly the exact same time and in the exact same developmental course across cultures, while others appear to come online in ways that are less tightly constrained in their ontogenetic timing (Callaghan et al., 2005; Liu et al., 2008). This suggests that such competencies may vary in the extent to which they are learned as opposed to innate. A theory that categorically denies the role of learning or the role of innately specified competencies is unlikely to account for these patterns of data. To that end, the theoretical framework that best fits the data and which informs subsequent work in this dissertation is likely some aggregate of the modular accounts proposed

by Leslie, Baron-Cohen, and Nichols and Stich. These, collectively, allow for canalized structures as well as learning, detail domain-specific modular structures, and are fairly maximalist in their characterization of the mindreading system. Critically, however, the position to be defended is one that does not uniquely rely on encapsulated functions, as proposed by Leslie and Baron-Cohen. While the components extracted from their models do retain these features, a number of others derive their functionality from interaction with other systems, like language, to which I now turn my attention. By reviewing the literature on the structural components and features of language and communication across human populations and across species, a theoretical understanding of language that bridges with mindreading is built.

Interactions between language, culture, and cognition in human beings

A longstanding debate in the anthropological literature has concerned itself with the relationship between language, culture, and cognition. While there is abundant evidence that language and culture vary quite significantly from one linguistic or cultural unit to the next, do these phenomena shape each other? And to what extent, if any, do they shape cognition? These questions have a lengthy history, with their antecedents identifiable in the work of some pre-Socratic Ancient Greek philosophers as early as the 4th century BCE (McComiskey, 2002). Nevertheless, the claim that language might shape thought did not receive a more recognizably modern form until the early 19th century when Wilhelm von Humboldt proposed, as part of a broader political, cultural, and intellectual project of German romantic nationalism, that language should be understood as the stuff of thought, the grammar of which represents the assumptions and beliefs of its corresponding nation (Verspoor & Pütz, 2000). Though scholars have debated whether von Humboldt's theory of language sought to justify or to mitigate colonialist views of national difference (Migge & Léglise, 2007; Said, 2016), he nevertheless argued that the dominance of German and English speakers over speakers of other languages was attributable to the grammatical perfection of the former languages over the latter (Verspoor & Pütz, 2000).

Thus, if language shaped thought, and some ways of thinking were “better” than others, then some languages might be “better” than others to the extent they facilitated or inhibited “better” ways of thinking. By the early 20th century, this notion had proliferated among American linguists and was used by some to argue for the eradication of Native American languages in the United States (Migge & Léglise, 2007; Seuren, 1998).

It was Franz Boas who first challenged the idea that some languages could be “better” or “worse” than others, arguing instead that all languages were equally modern, developed, and capable of expressing concepts. Notably, Boas appeared not to argue that the structure of a language could shape its speakers’ thoughts and thus their culture. Instead, he seemed to suggest that the thoughts of the members of a community could shape their culture. The structure of the language spoken by the community might then adapt to better encode the relevant cultural ideas and concepts (Boas, 1911). Nevertheless, Boas’ work was critical in dissociating the claim that language, culture, and thought may co-vary from the claim that such covariance entails the superiority or inferiority of a given language, culture, or way of thinking.

This development paved the way for Edward Sapir and Benjamin Lee Whorf to articulate the first truly modern accounts of linguistic relativity (Sapir, 1921, 1929; Sapir & Swadesh, 1946; Whorf, 1956). The arguments they presented were nuanced and subtle, if more metaphysical and less empirical in nature than later scholars understood them to be. Sapir, for example, argued that differences in grammar across languages corresponded to differences in the representation of reality. Accordingly, speakers of different languages ought to perceive reality differently (Sapir, 1929). Despite allowing for this possibility, Sapir recognized that the relations between language, culture, and thought were dissociable and non-deterministic. Speakers of a single language might not have a shared culture while speakers of multiple languages may participate in a broad monoculture (Sapir, 1921).

Similarly, Whorf argued that the grammar of a language was not merely a channel used by speakers to express their ideas, but a shaper of the ideas they might express. The structure of a speaker's language provided a lens through which to analyze their impressions of the world and carve it into meaningful categories. Per Whorf, "…the world is presented in a kaleidoscopic flux of impressions which has to be organized by our minds – and this means largely by the linguistic systems in our minds. We cut nature up, organize it into concepts, and ascribe significances as we do, largely because we are parties to an agreement to organize it in this way – an agreement that holds throughout our speech community and is codified in the patterns of our language..." (Whorf, 1956). Under Whorf's view, linguistic structure included grammar, but could be more broadly understood as referring to any of the patterns shared by word classes. Evidence for this interpretation of linguistic structure can be found in his analysis of cryptotypes, or grammatical categories that are not systematically morphologically marked by anything other than their shared implicit qualities and are "only definable negatively in terms of the restrictions they place on how morphemes can be combined" (Li, 1993; Whorf, 1956), such as the set of verbs that can take the prefix "un-". While most English speakers would agree that the verbs uncoil, untie, and unbutton are grammatically correct, many would likely disagree about the grammatically of verbs like unhate, unlook, or unsneeze. That these verbs differ in their ability to take the "un-" prefix is not indicated by anything other than speakers' sense that their application is incorrect in some cases and correct in others. Whorf posited that the shared semantic category to which these words belonged entailed something about "covering, enclosing, and surface-attaching meaning". This category, however, is otherwise unobservable in the structure of the language (Scholz et al., 2024).

Despite the predominant focus of Sapir and Whorf on grammar, and despite the relative temperance of their claims, their work was later mischaracterized by intellectual opponents and acolytes alike. Though the two never published together, and though neither Sapir nor Whorf

ever actually articulated a testable empirical hypothesis, the similarity of their ideas and their shared academic genealogy led subsequent scholars to retroactively lump their independent work together and to label their ideas the “Sapir-Whorf Hypothesis” (Hoijer, 1954). Brown and Lenneberg (1954), critics of linguistic relativism, went on to formulate a testable version of the “Whorf Hypothesis”, as they called it, which focused on the lexical codability of color categories. Across speakers of both English and Zuni, Brown and Lenneberg found that the lexical codability of a color was the strongest predictor of its recognition. This finding was taken to suggest the existence of a universal cognitive law relating a category’s codability to the underlying cognitive processes supporting recognition. In effect, because languages did not appear to vary in this relation, Brown and Lenneberg suggested that languages do not shape cognition – otherwise, the codability of a color may simply be one of many the linguistic qualities of color words that influence recognition. Because they did not observe a violation of the codability relation, Brown and Lenneberg concluded that the Whorf Hypothesis had not been supported by their data. It should be noted that Brown and Lenneberg’s treatment of Sapir and Whorf’s ideas was somewhat uncharitable, attributing to them the following tenets.

1) Different linguistic communities perceive and conceive reality in different ways.
2) The language spoken in a community helps to shape the cognitive structure of the individuals speaking that language.
3) Language is held to be causally related to cognitive structure.

This account has come to be known as “strong” linguistic relativism (Gumperz & Levinson, 1996; Penn, 2014) and it is what Brown and Lenneberg claimed to have disproven. However, the conclusions they drew appeared to minimize the fact that speakers of the Zuni language, which does not encode a lexical distinction between orange and yellow in the way English does, frequently failed to recognize these two colors correctly. Though this finding could be interpreted as illustrating that linguistic differences between Zuni and English caused a

cognitive difference in the recognizability of orange and yellow, Brown and Lenneberg maintained that they had found no data to support the Whorf hypothesis. They conceded, however, that though language does not *cause* cognitive structures, it may, "…be described as a molder of thought since speech is a patterned response that is learned only when the governing cognitive patterns have been grasped. It is also possible that the lexical structure of the speech he hears guides the infant in categorizing his environment" (R. W. Brown & Lenneberg, 1954). Accounts of this sort have come to be known as "weak" linguistic relativism (Gumperz & Levinson, 1996; Penn, 2014). Despite the early empirical support for and plausibility of its "weak" form, linguistic relativism fell out of fashion in subsequent decades.

Universalism became the predominant lens through which language structure was understood, influenced most famously by Chomsky and his theory of universal grammar (Chomsky, 1965). Universal grammar (UG) is posited to be an innate cognitive capacity for language acquisition with which all human beings are endowed. This capacity processes linguistic stimuli received in the course of language acquisition and imposes on it syntactic rules consistent with the principles of UG, however defined. In this way, UG creates structure to parse the incoming stream of speech to which children are exposed and to produce outgoing streams of speech that are meaningful to children's interlocutors. According to universalists, if human cognitive processes are universal, then there ought not to exist between-group differences in those cognitive processes. Though there are many readily observed differences between languages, if the fundamental structure of language could itself be understood as following from one among the many universal cognitive processes, then those differences must not have any meaningful causal effect on cognition. To attribute meaningful causal effects of these linguistic differences on cognition would be to erode the universalist position and as such, many scholars in this tradition, like Steven Pinker, became staunch anti-relativists (Pinker, 2003). Nevertheless,

recent decades have seen a reappraisal of linguistic relativity and a growing body of evidence that carves a path true to the tempered claims of Sapir and Whorf.

These more modern accounts appreciate the complexity of culture, language, and cognition and honor the fact that connections between them are likely to be as elaborate as the constructs from which they are derived. Additionally, they recognize that the extent of interaction between them likely relies on the cognitive domain at hand. Researchers in this space have taken the foundation laid by the Sapir-Whorf Hypothesis and have more carefully articulated the specific ways in which these three phenomena interact than did their predecessors, emphasizing that language may structure "habits of thinking" (Casasanto, 2015; Scholz et al., 2024). To the extent there exist multiple potential solutions to a representational problem, languages may vary in which ones they tend to encode. The brain is thus trained to solve that representation problem in that way because of how the language carves up the problem space. Though *all* solutions to *all* representational problems may be available to people the world over, the readiness with which any one solution is employed may vary cross-linguistically to the extent that it is encoded in the language more or less regularly. Claims of “weak” linguistic relativity such as these have proven to be powerful theoretical tools for studying the interaction of language, culture, and mind.

One major insight that cognitive science has contributed to this discussion is that language and cognition interact with each other in at least one significant way under a Representationalist view of cognition. That is, language maps representations and not objects as they exist out in the world. A consequence of this logic is that the entities to which language refers are not objective features of reality itself, but the subjective mental constructs used by speakers to represent it. As such, the boundaries drawn between words like “orange” and “yellow” do not exist independent of the representations held by speakers who use them. If such boundaries are not intrinsic features of reality, then from a place of first principles it is plausible

that people could carve it up in variable ways, perhaps even mapping concepts and representations upon which a culture has placed emphasis or importance. Note, however, that variability in the representations used to carve up reality and the language used to refer to them does not necessarily entail an inability to perceive or refer to other possible carvings. In fact, if boundaries encoded by language are not intrinsic properties of reality itself, it seems more likely that speakers could hold and refer to representations of reality specified in potentially many ways. Nevertheless, there are reasons to be skeptical of these claims.

To the extent it is important for an organism to get a certain representation "right", there ought not to be meaningful variance between individuals or cultural groups in the ability to have and employ that representation. Such representations ought to have been selected for over evolutionary history to be independent of and impermeable to the influence of other processes, cognitive, external, or otherwise. Having painted such a picture, it is clear that a more nuanced treatment of the ways in which language and culture shape cognition is needed. I now turn towards some of the evidence in favor of such linguistic and cultural effects on perception and representation. After, I review the evidence against them.

Evidence for linguistic and cultural effects on cognition

To date, several prominent authors have contributed data that paint a rich and compelling picture regarding the permeability of higher-order cognitive processes concerned with perceptual representations to the influence of language and culture. Though some have argued that perceptual processes and perceptual representations may themselves be permeable to language and culture (Balcetis, 2016; Collins & Olson, 2014; Dunning & Balcetis, 2013; Goldstone et al., 2015; Lupyan, 2012), others have argued compellingly that these reported effects are more meaningfully understood as influences on cognition *about* perception (Firestone & Scholl, 2016). Regardless of whether these effects manifest in perception itself or in cognition about perception, they may still be domain-specific – that is, some representational

domains may be more reliably shaped by culture and language than others. Many of the key findings in favor of linguistic and cultural effects on cognition, for example, have either emphasized cross-cultural differences in the way that perceptual spectra are carved in a language or have focused on the kinds of representational and perceptual phenomena for which there may have been strong cultural, but not natural selection. Early work by Berlin and Kay (1969) showed that across societies, not all languages divide up the electromagnetic spectrum into the same set of basic colors. As universalists, Berlin and Kay argued that there exist a universal set of 11 basic color terms and that the number of basic color terms a language had (ranging from 2 to 11) could reliably predict the colors to which those terms referred. Thus, all languages have terms for black and white. If a language has three basic color terms, then it has a term for red. If a language has four basic color terms, then it has a term for green or yellow, and so forth (Kay & Regier, 2003, 2006; Regier & Kay, 2009). Nevertheless, the fact that the languages Berlin and Kay reviewed varied in the number of basic color terms supports a kind of relativist position. Specifically, languages appear to inform the cognitive processes that are capable of drawing (or not drawing) categorical boundaries on perceptual representations and that this ability is independent of the fitness consequences of being able to perceive a particular color. Evidence for these kinds of effects is not limited to the domain of color perception, with such effects evident in the domains of spatial perception and navigation, moral decision-making, and olfaction (Fausey et al., 2009; Fausey & Boroditsky, 2008, 2010, 2011; Giannakopoulou et al., 2013; Haun et al., 2011; Hevia et al., 2014; Majid et al., 2018; Munnich et al., 2001; Tajima & Duffield, 2012; Wnuk & Majid, 2014; Wolff & Holmes, 2011).

Spatial perception and navigation

Research on spatial perception and navigation provide compelling evidence for the influence of language on thought (Giannakopoulou et al., 2013; Haun et al., 2011; Hevia et al., 2014; Munnich et al., 2001). While spatial concepts like left and right might appear to an English

speaker as objective and self-evident ways to characterize space, one need not look beyond English to find evidence that paints a more complex picture. The words port and starboard, for example, differ from the words left and right in terms of the points of reference to which they are anchored. Whereas port and starboard are fixed to external points of reference, left and right are fix to individuals' perspectives. While a life vessel will be portside no matter my perspective, whether it is to the left of me will vary as a function of my location in space.

Two important points are borne out here. The first is that different linguistic systems for realizing spatial navigation entail different cognitive demands. A naval officer will need to maintain an active representation of their own orientation with regard to the ship's boundaries, while civilians on a cruise ship may face no such demands. When instructed to board a life vessel on the port side of the ship, officers and civilians might reasonably be expected to differ in the efficiency with which they process such instructions. These differences can be understood as "habits of thought", or differences in the regularity, and thus, efficiency, with which a particular representational format is called upon. Both civilians and naval officers can come to be experts in the use of the terms port and starboard, but it is only those who use the terminological system and thus regularly employ the representations they index that use them efficiently. Indeed, these effects have been documented in a number of languages whose terms for spatial navigation are explicitly geocentric, as opposed to predominantly egocentric languages like English (Burgess, 2006; Dasen & Mishra, 2010)

The second point is that there are no given or exogenous concepts out in the world for solving linguistic coordination problems about space. The only parameter that matters is whether or not all users of a given terminological system agree about their referents. While certain conceptual solutions may present themselves more readily than others in virtue of inherent differences in their cognitive salience, a given set of spatial terms is no more natural or correct than any other. In theory, then, there are potentially limitless systems of spatial

reference. Indeed, many languages have been shown to use landmarks like mountains and the flow of rivers to anchor spatial reference (Giannakopoulou et al., 2013).

Moral decision-making

More recently, a large body of literature has shown that many of the decisions and judgments we make about others are tied intimately to our social and cultural contexts. Thus, a given action undertaken by an individual may have differential fitness consequences across cultures and societies. The way in which those actions are represented and thus judged appears to be predicated at least partially upon how it is encoded linguistically. (Fausey et al., 2009; Fausey & Boroditsky, 2008; Haun et al., 2011; Tajima & Duffield, 2012; Wolff & Holmes, 2011). For example, given that human beings assign moral judgments to actions based on the extent of their intentionality (although see (Barrett et al., 2016) for a discussion), and that one's intentions are generally not perceptually available, language may shape the way in which moral judgments are made about the actions of others (Fausey & Boroditsky, 2010). Additionally, certain representational *types* may themselves only be meaningful in a particular cultural environment. The constitutive features of those types may be perceptible everywhere, but their collective occurrence as distinct entities with causal behavioral power likely relies entirely on the historical, cultural, and linguistic circumstances that preempt the utility of such entities in the first place. Work on social identity in Chinese and English has shown such evidence (Hoffman et al., 1986), suggesting that language and culture may themselves structure and produce representations that otherwise have no salient causal power.

Olfaction

Olfaction constitutes another domain in which the effect of language on the cognitive processes that categorize perceptual input has been found. Majid and Burenhult (2014) found that speakers of a language with a rich olfactory lexicon (Jahai) could name odors as readily as colors, in contrast to speakers of a language without the same richness of odor terms (English).

These results suggest, in a way similar to those of Brown and Lenneberg (1954), that linguistic encoding of perceptual phenomena may facilitate the ease with which they are recognized. Extending this perspective, Majid et al. (2018) examined sensory codability across 20 languages to determine whether there exists a universal cross-linguistic hierarchy of the senses with respect to how readily they are accessed by consciousness and available to linguistic description. Critically, the specific sensory modalities that were systematically linguistically encoded, as well as they ways in which they were encoded, varied across languages. The authors posited the tendency to code more effectively for a given domain may be attributable to preoccupations with that sense in a particular cultural context, a fact that suggests the flexibility of higher-order cognitive processes when categorizing perceptual stimuli.

Evidence against linguistic and cultural effects on cognition

In contrast to the findings reported above, a number of other scholars in the fields of linguistics and cognitive science have generated a body of evidence which has been taken to disconfirm the claims of linguistic relativists (Chomsky, 1965; Goddard & Wierzbicka, 1994, 2002; J. H. Greenberg, 1963; Heine, 1997; Pinker, 2003; Pinker & Bloom, 1990; Rosch et al., 1976; Wierzbicka, 1972, 1992, 1996). While it is plainly the case that languages differ in their phonological inventories, their morphologies, and their syntax, these authors present arguments to suggest both the universality of the cognitive processes that support the reliable development of language, irrespective of its particular form, and the universality of the cognitive processes with which languages interact. If both of these classes of cognitive processes are in fact universal, it is implied that whatever variation can be observed across languages must not exert differential effects on the cognition their speakers.

Data consistent universality in the cognitive processes supporting the reliable development of language have emerged predominantly in the wake of Chomsky's theory of Universal Grammar, or UG (Chomsky, 1965; J. H. Greenberg, 1963; Heine, 1997; Pinker, 2003;

Pinker & Bloom, 1990). The basic precept of UG is that there are innate constraints on what the grammar of a possible human language could be. These innate constraints thus provide pre-linguistic children with tools for parsing the incoming stream of linguistic stimuli to which they are exposed. Though the grammatical particulars of any two languages may differ significantly, they represent equivalent "solutions" to the functional problem language is meant to address. Though the grammatical possibility space is mapped in different ways, they nevertheless fulfill the same basic function and achieve the same basic outcomes. By analogy, suppose I want to make a map of Los Angeles County. Regardless of whether I want to make a road map, a geological map, or a topographic map, there is a substantive constraint that applies to all three. Namely, they must meaningfully depict the spatial arrangement of the county in a way that accurately represents distance. The scale used and the particular features represented on the map are free to vary, but they all equally well solve the problem of "mapping Los Angeles County", or meaningfully depicting its spatial arrangement. Data to support the theory of UG has been drawn from the creolization of pidgin languages by native speakers born into such contexts of language contact, demonstrating the emergence of consistent grammatical structure out of a non-structured system of communication (Bickerton, 1984). Similarly, it has been claimed that certain grammatical or syntactic properties must themselves be universal (J. H. Greenberg, 1963).

While supporters of UG have pointed to the universality of the developmental and structural properties of the world's languages as evidence against linguistic relativism, others have instead focused on the universality of the meaning conveyed by their lexica (Fodor, 1975; Goddard & Wierzbicka, 1994, 2002; Rosch et al., 1976; Wierzbicka, 1972, 1992, 1996). That is, some have argued that in order for human language to function, it must make communicable some minimal set of essential semantic concepts regardless of their grammaticalization. As such, these "semantic primitives" should be present in all of the world's languages. Researchers

across a broad range of fields have articulated repeatedly that human brains are equipped with the same cognitive abilities everywhere and it is through them that individuals arrive at their conceptions and representations of the world. Because language may only describe the world as people conceive of it, and because people everywhere are endowed with the same cognitive tools, it stands to reason that at least some of the ways in which people construe the world may be the same across populations. The universality of verbs and nouns as grammatical classes (Pinker & Bloom, 1990), for example, may be a consequence of the fact that the brain parses events as distinct from objects, broadly speaking. At a less abstract level, all languages may have a word or words for mother given the cognitive mechanisms with which mammals are equipped to identify one's primary caretaker. To the extent such mechanisms are themselves universal, so too may be their influence on the lexicons of the worlds' languages. Under a strict interpretation of this view, substantive cross-linguistic variation in speakers' conceptions and representations of the world and its features ought not to be observed. In effect, we ought not to view meaningful variation across languages in the representation of these semantic primitives in their lexica (Fodor, 1975; Goddard & Wierzbicka, 1994; Wierzbicka, 1972, 1992, 1996).

In accordance with such a prediction, there are a number of domains in which effects of language and culture on cognition have not been found. Despite promising early findings on the role of color perception and the effects of language, later research showed that although languages vary significantly in the culturally evolved suite of color terms they have, a few notable patterns could be observed (Berlin & Kay, 1969). Although cultural groups vary in their color terms, those they have seem to be a function of the size of the lexical inventory for colors. Languages with just two terms for color tend to encode black and white, while those with three encode black, white, and red. This pattern is robust and suggests that their perceptual qualities may be equally salient across cultural milieus, constituting a kind of cultural attractor (Sperber, 1996). This notion is distinct from that of semantic primes or primitives, which suggests that

every language shares a core vocabulary of concepts (Wierzbicka, 1972, 1992, 1996). Because it has been claimed that every language has terms for black and white, these might be among the set of semantic primes or primitives shared by every language. However, as additional color terms are added to the lexicon, cognitive or perceptual biases may also impose a universal order according to which richer ranges of color concepts are built out. Additional evidence of this claim has been found with neural imaging studies, showing that the same regions of the brain tend to react to the same color stimuli despite differences in languages' lexical inventories for color (Bornstein, 2006; Bornstein et al., 1976).

Beyond color perception, there are many other domains in which linguistic relativity has not been found. Given a powerful history of selection across phylogeny, it is likely to be the case that all human sensory systems' functions are language-independent. Although recent studies have shown the effects of linguistic relativity in the domain of olfaction (Cain et al., 1994; Lehrner et al., 1999; Majid et al., 2018; Oleszkiewicz et al., 2016; Wnuk & Majid, 2014), there remains the fact that, barring mutations in olfactory bulb chemoreceptors, all human beings everywhere are equipped with a functionally identical capacity to detect odors. While language might upregulate attention allocated to scent, it is unlikely to have exerted top-down control on the breadth of detectable scents. A similar conclusion ought to be true of audition. An ear can be "tuned", but all listeners' auditory cortices are processing the same soundwaves. As a general principle, the more concretely a given target can be shown to exist independently of its representations, the less likely it subject to the effects of linguistic relativity. In contrast, the greater the extent to which a given target is "in the mind", the more sensitive its contours may be to them. To a certain extent, the presence or absence of these effects will also depend upon the granularity of any such analysis. Consider, for example, the detection of biological motion. While there is robust evidence for language-independent psychological mechanisms to detect biological motion (Castelli et al., 2000; Simion et al., 2008), the specific elements to which

individuals attend may be conditioned on the language spoken. Fausey and Boroditsky have shown that attributions of agency to an actor's actions can be up- or down-regulated as a function of the language used (Fausey & Boroditsky, 2008). It is thus important to specify the features of representations sensitive to language.

Conclusion

Collectively, these data and theoretical positions illustrate that there exist at least some cognitive phenomena which are sensitive to and structured by language, though the extent of such effects is nuanced and can be used to support both universalist and relativist positions. Additionally, there exist myriad ways that such effects on cognition may be instantiated. With these arguments at hand, I turn now to the question of whether mental state talk varies cross-linguistically, and if so, whether it bears any relation to variation in mindreading.

Does mental state talk vary across languages and if so, does it matter?

Mindreading and language are intimately interconnected, both in terms of how mindreading undergirds the capacity for language itself and how human communication very often serves to influence the content of others' minds (if not to communicate about it outright). Each of these constructs alone are critical to making predictions about how others will act, but the value of their linkage to these ends cannot be understated. Unlike other species, people can tell you what their goals and intentions are, collapsing uncertainty about the targets of an individual's actions and reducing the computational load faced by the mindreading system. However, languages vary across a tremendous number of parameters, which is to say nothing of the variation that exists both between individual speakers and between different communities of a single language. An implication, then, is that mindreading might vary across individuals, cultures, and languages (Goddard, 2010; Lillard, 1998). In the following paragraphs, several areas of interaction between language and mindreading will be outlined to determine if the

current evidence supports the notion that mental state talk varies cross-linguistically and if so, whether it correlates with variation in the mindreading capacity.

Pragmatics and norms

Distinct linguistic components can contribute to sentence meaning (e.g., morphological, syntactic, and semantic), but pragmatics examines situated, contextual meaning (sometimes called speaker meaning). A significant portion of the meaning of the utterance "uh, yeah, sure" will depend on the context in which it occurs. While given in the actual semantic meaning of the utterance itself is some degree of positive affirmation, how that affirmation ought to be interpreted will be conditioned heavily on who is saying it, to whom it was directed, the audience, where it happened, the intonation, and the discursive milieu to name just a few of the pertinent factors. "Uh, yeah, sure," is a fine answer to follow the question "Can I borrow a dollar?" but a deeply troubling answer to "Will you marry me?". Languages afford their speakers many tools to conceal or convey their mental states. While this flexibility presents potential challenges to language processing, it poses an especially potent one to the mindreading system. Indeed, several theorists have suggested a uniquely rich role of mindreading in processing the pragmatics of communicative acts (H. Clark, 1996; Scott-Phillips, 2014, p.; Sperber & Wilson, 2001). Pragmatics as a field is often defined in relation to semantics. Where semantics refers to the meaning of a word or sentence per se, pragmatics refers to the effect context exercises on the meaning of language. Thus, one way mindreading interacts with language is by disambiguating the meaning of tokens of known communicative structures in situ, as well as interpreting novel communicative structures in a rapid, online manner (Misyak et al., 2016). For example, suppose my partner asks me to, “get the red thing.” Though I am familiar with each of these words and their semantic mapping, I rely on my representation of her beliefs, desires, emotions, percepts, and intentions to disambiguate the specific item to which the token “thing” refers. Similarly, suppose now my partner asks me to “get the red wug.”

Though I am unfamiliar with the word “wug” and its semantic mapping, I can draw on both my own knowledge (or lack thereof) and my representation of her beliefs, desires, emotions, percepts, and intentions to infer the plausible referent of her request.

Crucially, it is important to be clear about what constitutes “context.” While features of language itself, like intonation in the example above, certainly count, the range of phenomena that may condition semantic meaning is broad. One such phenomenon is the set of communicative norms a given language community uses. For example, consider the difference between high-context and low-context cultures. High- and low-context cultures represent ends of a continuum with respect to the explicitness and context-dependence of communicative exchange (E. T. Hall, 1973). High-context cultures often exhibit less direct verbal and non-verbal communication, with more meaning read into these more indirect messages. High-context cultures, in contrast to low-context cultures, may thus operate such that all members have onboarded an extensive suite of norms and their associated social meanings. Where violations of such norms occur, all members of a high-context culture may read the same meaning into the implicit and indirect violation. In turn, there may be less explicit communication about the intent behind the violation. If everyone knows it is a violation, including the violator, then it must have been intentional. As such, speech acts pertaining to the mental states of the violator, such as their intent, may manifest less readily than in lower-context cultures.

To the extent it can be assumed one’s interlocutors share a similar conception of the world, there may be less reason to speak about the nature of that conception. Additionally, these dynamics may not apply uniformly to all categories of mental states. Some kinds of mindreading may be offloaded into a shared conception of the world while others remain free to be adjudicated through speech (Robbins & Rumsey, 2008). For example, a component of British national identity, at least historically, has been “to keep a stiff upper lip”, or to minimize the expression of emotion in the face of adversity (Storry et al., 2002). It may thus be reasonable to

expect the production of fewer speech acts pertaining to one's emotional experiences. In contrast, if there are elements of British identity concerned with the expression of epistemic mental states like belief or intention they are not nearly as widely known as those components concerned with limiting emotional expression. Nevertheless, it is at least plausible to think that there are no such values placed on the suppression of such mental states, and that there may in fact be value placed on expression of beliefs and intentions. The adoption of common law following the Norman conquest of England in 1066 and the subsequent incorporation of the standard of *mens rea* from canonical law, with its focus on intent and knowledge, point toward what is potentially a cultural hyper cognizance of such mental states (Noyes, 1944). Thus, speech about belief and intent may occur with greater frequency than speech about emotions.

Other normative phenomena that may be of relevance to mindreading include the quality and quantity of child-directed speech. Across the world, there exists substantial variation in the extent to which child-directed speech exhibits patterns of "child-raising", or interacting with children as if they were fully competent interlocutors, and "caretaker-lowering", or catering interactions to a child's interlocutory competence. Embedded within this framing of child-directed speech is also the extent to which children receive direct communicative engagement as opposed to indirect absorption of interaction occurring around them (Akhtar & Gernsbacher, 2007; Perner et al., 1994; Ruffman et al., 2002). Across such contexts, the quantity of mental state talk to which children are exposed, as well as the quantity of opportunities within which to develop mastery of the concepts indexed by such speech, may be variable. Even if the production of mental state talk between adults varies minimally across cultural or linguistic contexts, there may nevertheless be variation in the production of mental state talk by adults to children, yielding differences in exposure and subsequent mastery.

Across all of the phenomena discussed here, one possible consequence could be the presence of genuine differences in mental state concepts across cultures. Alternatively, cross-

linguistic normative and pragmatic differences may represent variation in the priority and attention given to particular mental state concepts borne by individuals within each culture but drawn from a universally shared suite of mental state concepts. In that way, culture may be structuring not the size or set of conceptual tools, but the ones that are more readily drawn upon and realized in day-to-day social contexts and interactions. Moreover, norms of communication may structure the kinds of pragmatic inferences that are drawn from a given utterance or speech act. Whereas certain kinds of utterances may reliably indicate something about the mental states a speaker intends to communicate, such pragmatic inferences may be unwarranted in another cultural context. Similarly, norms of communication may just increase or decrease individuals' exposure to mental state terms, a possibility that is implied by the arguments which have been made about the existence of mental opacity cultures (Robbins & Rumsey, 2008). If exposure is a factor that determines the rate at which mindreading matures, and communicative norms can shape the relative frequency of exposure to those terms, then there may be reciprocal feedback between these processes.

Semantics and the lexicon

It is not the quantity of overall language exposure that predicts children's performance on the False Belief task, but the quantity of mental state verbs to which they are exposed (Bretherton & Beeghly, 1982; Brooks & Meltzoff, 2015; J. R. Brown et al., 1996; Ruffman et al., 2002). Mental state verbs, as lexical items, may be unique in their ability to track mental structures given their ability to take whole independent clauses, or phrases that can stand alone as sentences, as their grammatical objects. In grammar, any word, phrase, or clause that is required to complete the meaning of a sentence or a part of a sentence is called a complement. As mental state verbs require both subjects and objects, independent clauses that serve as the object of a mental state verb are sometimes called sentential complements. In English, independent clauses can be made into sentential complements through the use of the

complementizer “that”. As an example, the sentence, “Anna believes that John is the tallest student in class” contains the independent clause “John is the tallest student in class” as a sentential complement of the verb “believe”. This grammatical property has been argued by some (de Villiers & Pyers, 2002; Durrleman et al., 2019; Gleitman, 1990) to constitute a linguistic parallel of the epistemological properties of mental state representations that allow individuals to understand others’ false beliefs. That is, a sentential complement can be false even though the sentence is true overall in much the same what that the content of another person’s belief may be false, though it is true they hold that belief. Though the moon may not be made of cheese, John may well believe that it is. The uniformity with which mental state verbs support constructures of this kind may play an explanatory role in their relationship to mindreading development by providing a linguistic infrastructure through which to learn the truth conditions of such nested statements and how they relate to the unique contents of others’ minds (Gleitman, 1990). Moreover, the way mental state concepts are lexicalized may be universal (Goddard, 2010).

Syntax

As mentioned above, mental state verbs have a predictable syntactic structure that allows them to take sentential complements. Practically all English-language mental state verbs are of this kind. Such verbs are special because to "do" them is to engage in a type of action whose only role is to hold a state of affairs that need not be true, accessible to perception, or endorsed by the actor. In this way, children need only track both the circumstances in which the word is occurring as well as other contextual information, like the syntactic context and the statistical occurrence of such verb forms to bootstrap understanding of certain otherwise opaque concepts (Gleitman, 1990). In the case of sentential complement verbs, the child can learn two things from them – the first is that language can be used to make observable some kinds of otherwise unobservable entities, giving them perceptible form. While the concept or

representation indexed by the sentential complement of a mental state verb may not refer to anything real, the fact that it can be expressed linguistically means that it can nevertheless have actual causal properties. A constrained example of what I mean by actual causal properties can be seen in the pretend play of children (Leslie, 1987; Nichols & Stich, 2003). If one child says to another, “Let’s pretend that the floor is made of lava”, they may subsequently avoid touching the floor, jumping on furniture and screaming when one does accidentally make contact with it. Though the sentential complement “the floor is made of lava” does not refer to something real, its expression nevertheless exerts causal influence on the children’s subsequent behavior.

The second thing the child can learn from the grammatical properties of mental state verbs is that some features of their experience, which may be eminently observable from their own perspective, must be communicated through the same syntactic structures to be observable to others. Beyond syntax's role in the development of mindreading, some evidence suggests that syntax may make contact with mindreading in the adult speech of some languages. Specifically, some languages have what are known as obligate morphological evidentials, or grammatically necessary markers of the evidentiary basis for a given statement, such as through direct knowledge, inference, or hearsay (Aikhenvald, 2004). Speakers of languages with obligate evidential structures may facilitate the extent to which their speakers scrutinize the claims made others, a phenomenon known more broadly as epistemic vigilance. The tendency to track the source of information claimed by others may entail, as a consequence of which, tracking the knowledge states of others (Aikhenvald & Dixon, 2003; Sperber et al., 2010; Tosun et al., 2013).

Impacts of language on mindreading

While structural features of a language may place constraints on the kinds of output a mindreading system produces, so too do the ways a language carves up conceptual and perceptual space. This notion suggests that language defines the boundaries on what is

otherwise continuous conceptual and perceptual space. These boundaries can structure one's conscious experience by facilitating the retrieval of certain memories over others, making certain aspects of the environment more or less salient, and so on. Critically, languages vary in where such boundaries are drawn in virtue of the fact that they are merely impositions on the continuous conceptual and perceptual space. As such, the conscious experience of a scene by speakers of two different languages may differ in systematic ways despite the commonality of the stimulus to which they are exposed (Boroditsky, 2011; Casasanto, 2015; Wolff & Holmes, 2011). As an example, consider a language in which terms for animals are specified at the level of genus or family versus a language in which animal terms are specified at the level of species. In the first language, a single term could be used to refer equally well to dogs, wolves, jackals, and coyotes, effectively marking the differences between these species as insufficiently meaningful to differentiate between. In the second language, distinct terms would be needed for each, suggesting that those differences constitute meaningful boundaries worth tracking as distinct. An image of a dog, a wolf, a jackal, and a coyote presented to speakers of each of these two languages is perceptually identically but what is represented by this percept may nevertheless be very different. Findings of this kind have been observed in the domains of number, color, smell, and perhaps of greatest importance to the current discussion, emotions and mental states (Cheung et al., 2009; Jackson et al., 2019; Kay & Regier, 2006; Saalbach & Imai, 2011; Wnuk & Majid, 2014). Beyond abstract conceptual and perceptual domains conditioning mindreading system inputs, some mindreading phenomena appear themselves to be affected by linguistic factors, including person cognition (Hoffman et al., 1986), attributions of intentionality (Fausey & Boroditsky, 2008; Hargreaves, 2005), and memory for persons (Fausey et al., 2009; Fausey & Boroditsky, 2011).

Conclusion

Taken together, these studies provide evidence to suggest that some components of language may exhibit variation with respect to talk about the mind whereas others may manifest more uniformly across cultural and linguistic environments. While there is also evidence to suggest that at least some elements of the mindreading capacity exhibit cross-cultural variation, there are very few studies to date that permit strong conclusions to be drawn about the influence of such variation on mindreading cognition. Consequently, there is an urgent need for both systematic studies of mental state talk phenomena to determine, definitively, which truly exhibit cross-linguistic variation, as well as studies that can illustrate a relationship between variation in a given element of mental state talk and variation in mindreading.

A plan to address outstanding questions

Having sketched some of the theoretical and empirical landscape, several tensions now present themselves. Given past studies (Gleitman, 1990; Liu et al., 2008; K. Milligan et al., 2007; Papafragou et al., 2007; Perez-Zapata et al., 2016; Robbins & Rumsey, 2008; Wellman & Liu, 2004), one might conclude that there really is a relationship between an individual's exposure to mental state talk and their ability to read the minds of others, in which case it must be admitted that there might be differences in mindreading ability across cultural groups and language communities. This seems like an unsavory conclusion to commit one's self to, especially given that there is no evidence to suggest either cross-population differences in average sociocognitive ability or cross-population differences in the presence and effect of distinct selection pressures on social cognition (but see Bradford et al., 2018). Moreover, why would evolution favor an adaptation for language that permits the wholesale absence of features that matter for the development of mindreading if mindreading is as critical an adaptation as has been argued in the literature? If mental state language really does matter for mindreading

development, one might expect evolution to have placed tighter constraints on the production of mental state language so as to guarantee its reliable development across variable cultural environments.

Taken together, these data suggest at least one of the following possibilities. The mindreading system may have evolved to take language as an indexical input of the averaged mind-mindedness of potential interlocutors in the social environment, thereby affording adaptive plasticity in the allocation of resources to the development of mindreading. While there is some evidence for the influence of social ecology on language structure (Dale & Lupyan, 2012; Lupyan & Dale, 2010; Nettle, 2012), there is to date no evidence suggesting such effects with specific respect to linguistically encoded semantic domains, like mental-state language. Far more likely to be the case is either the available data have suggested greater differences in the production of mental state talk than actually exist across cultures and languages, or the relation of an individual's mindreading ability to their production of mental-state talk has been overstated.

Despite the clear importance of a resolution to this conflict for our understanding of the mindreading system, the language system, and the evolution of their support structures, two fundamental questions have remained unaddressed. Do language communities actually vary in their production of mental-state talk? And across cultures, is an individual's production of mental state talk a meaningful predictor of their mindreading ability? This dissertation aims to address exactly these questions. Answers to these two questions can inform hypotheses of relevance to future research. For example, given extant data suggesting the role of mental-state talk on mindreading development among English speaking-children, would this pattern hold cross-linguistically? If not, why?

Other future research questions this dissertation can speak to are as follows. What function, if any, does it serve to encode mental states linguistically? Can this function be

achieved through other means? Does the architecture of the mindreading system place upper and lower bounds on the frequency with which mental states are encoded in language? Does the mindreading system have canalized, bottom-up inputs to the linguistic system which are then filtered out or left in depending on cultural or linguistic practices? Or do linguistic and cultural norms, in conjunction with top-down mindreading processes, construct utterances with mental state talk as is deemed relevant? Does variation in linguistic and cultural practices surrounding mental states exert feed forward influence on the kinds of mindreading people tend to do across cultures and languages? Even if no such relationship exists between them, why do so many languages have words or morphemes allowing speakers to encode mental states linguistically?

While all of these questions are critical to understanding the evolution of language and the mindreading system, this dissertation aims only to answer the first two, restated here.

1) Do language communities actually vary in their realization of mental-state talk?
2) Across cultures, does an individual's production of mental-state talk predict mindreading ability?

In the following pages, I will describe a novel and cross-linguistically generalizable methodology for the production of systematic and standardized corpora of speech samples about the minds of others. These corpora will contain measures of participants' mindreading abilities, samples of participant's elicited speech about a controlled set of video stimuli, and a variety of demographic measures. American English, Moroccan Arabic, and Mandarin Chinese participants were recruited from the United States, Morocco, and China, respectively, to generate these data. Because few prior studies have actually measured the frequency of mental-state talk across distinct language communities, whether it varies at all is unknown at present, to say nothing of the factors responsible for such variation if indeed it exists. Therefore, I sought participants from these populations owing in part to the many dimensions of difference

between them which have been posited to effect variation in both mindreading ability and in the production of mental-state talk. By doing so, I aimed to maximize the likelihood of observing cross-linguistic variation in mental-state talk. If no variation was found between these samples, I could thus more confidently conclude the generalizability of the finding.

These data will permit quantification of the difference (or lack thereof) in the following directly observed measures.

1) Individual and cultural level differences in the production of mental state language given observation of the same set of stimuli
2) Individual and cultural level differences in mental state language given varying contexts and situations
3) Individual and cultural level differences in mindreading

Beyond just quantifying group- and individual-level differences in the measured variables, the relationships between these variables will be assessed using these data. Specifically, these data will permit a number of analyses to be performed that will assess the contribution of mindreading ability to the production of mental state language in elicited contexts across cultures. Three distinct sets of analyses will be described. The analyses within each set have been grouped as follows. In the first set of analyses, speech samples are processed and coded using a set of verbs drawn from the literature to determine whether speakers of these three languages vary in the average frequency of their mental state talk. In the second set of analyses, the same corpus of speech samples is processed and coded using a different set of terms. In this case, native speakers of each target language were tasked with identifying all of the words from the corpus that could plausibly constitute lexical references to third-party mental states in order to improve upon and expand the narrow set of terms captured in the first coding scheme. Finally, these data are used in combination with participant performance on a measure of mindreading to determine whether these two variables correlate. Taken together, these

studies represent the first systematic and quantitative study of the interrelationship between mindreading and mental state language, across three dramatically different language communities. The results of these studies will speak to the presence or absence of variation in mental state talk between individuals and across language communities as well as provide preliminary evidence about the extent to which these factors covary cross-linguistically.

Chapter 2: General Methodology

Introduction

Having provided a broad map of the research landscape and the core questions to be addressed in this dissertation, data capable of answering them must meet the following criteria:

1) Sampled from distinct populations
2) Validity of tools and measures does not vary across populations
3) Capture naturally produced spoken language

In this chapter, the population of participants sampled, the measurement tools, the procedures dictating their use, and some of the methods used to clean and process the data are described. Those methods and procedures unique to each of the three planned sets of analyses are detailed in the chapters to which they correspond. Nevertheless, a high-level view of the general methodology is as follows. First, video stimuli of social scenes with rich character motivations were created de novo. These stimuli were then shown to speakers of three languages – Arabic, English, and Mandarin Chinese – recruited from three countries – Morocco, the United States, and the People's Republic of China. After viewing a given stimulus, participants provided descriptions of what they had seen. Participants watched a total of 9 videos, provided demographic data about themselves, and completed a commonly used measure of mindreading ability – the Reading the Mind in the Eyes Test (Baron-Cohen et al., 2001). Participant descriptions were recorded, transcribed, and coded to quantify participants' mental state talk about the characters in the videos. These values were then modeled as a function of participant demographic variables and their mindreading ability to determine whether they correlated with participants' mental state talk.

Methods

Participants

Participants were recruited from collaborating field sites and institutions in China, Morocco, and the United States as part of the Geography of Philosophy Project (GPP), a research initiative funded by the John Templeton Foundation from 2018 to 2021 to explore universality and diversity in fundamental philosophical concepts. Broadly, the goals of the GPP are "to advance what is known about the extent to which three fundamental philosophical concepts – knowledge, understanding, and wisdom – are shared across religions and cultures" and "to create a new, multi-cultural research community focused on studying important philosophical concepts using the tools and insights of a wide variety of disciplines including philosophy, anthropology, linguistics, psychology, neuroscience, and cultural studies" (Geography of Philosophy Project, 2017).

Participants were deemed ineligible if any of the following criteria held: 1) they were below the age of 18, 2) they were not fluent L1 speakers of the target language, and 3) they did not reside full-time and long-term in the country from which they were recruited. That is, participants were not filtered out as a function of citizenship or legal status in the country from which they were recruited. Rather, they were deemed ineligible if there was sufficient evidence to suggest that their cultural and linguistic experiences were informed by the broader culture of a country other than the three from which the samples were drawn. Inclusion or exclusion according to this criterion was determined by participants' self-reported nationality and country of residence. Target languages were Mandarin Chinese for participants recruited in China, Moroccan Arabic in Morocco, and American English in the United States.

English-speaking participants in the United States were all students attending the University of California, Los Angeles. They were recruited using the UCLA Department of Communication Subject Pool and were awarded research credits for their time. Arabic-speaking

participants in Morocco from were sampled from two distinct populations. First, Arabic-speaking students attending the International University of Rabat were recruited and participated on a voluntary basis. Second, Arabic-speaking members of the public living in Rabat were recruited and participated on a voluntary basis as well. Mandarin-speaking participants in the People's Republic of China were students attending Xiamen University. They were recruited by snowball sampling with an initial pool of participants drawn from the philosophy department and subsequent pools drawn from referrals provided by the initial pool. Participants were invited on a voluntary basis.

No fewer than 40 participants were recruited from each population (and sub-population), yielding a total expected n = 160. Upon completion of data collection, a total of 191 subjects from the United States (n=56), China (n=53) and Morocco (n=82) had participated in the study. After the removal of participants deemed ineligible according to the criteria detailed above, a total of 177 participants remained, with 1 Chinese participant and 13 US participants removed. All participants provided a complete set of 9 video vignette descriptions, except for three participants who were each missing a single description, yielding a total set of 1589 descriptions. To the greatest extent possible, equal numbers of men and women were recruited as participants from each population. Additionally, participants were recruited so as to minimize differences in mean age across the sites, where possible and pertinent. These demographic variables were matched in order to account for the well-documented effects of sex and age on linguistic practices and mindreading ability. These are, respectively, that female individuals score more highly on measures of mindreading relative to male individuals and that performance increases through adolescence followed by a shallow decline across adulthood (Baron-Cohen & Wheelwright, 2004; D. M. Greenberg et al., 2023; Haselton & Buss, 2000; Newman et al., 2008; Prewitt-Freilino et al., 2012). These samples represent strong candidates for the observation of variance in mental-state talk due to their significant differences along a

number of dimensions that plausibly pertain to its production, including variation in the religious, ethnic, linguistic, and family demographic composition of each sample. The relevance of these dimensions to mindreading is detailed in the next sections. It warrants mention that these factors do not represent an exhaustive account of the differences between these samples, Instead, they represent those dimensions for which there is theoretical or empirical work suggesting their role in mental state talk. Therefore, if it is found that these populations do not differ in mindreading ability or in the production of mental state talk, there may be other variables pertinent to these capacities along which these three populations do not, in fact, vary.

Religion

It has been suggested that mindreading ability and religiosity may be related such that more religious individuals tend to attribute intentionality to a greater range of inanimate entities than less religious individuals. To the extent that the attribution of intentionality is a component of mindreading, religiosity as a trait may interact meaningfully with the mindreading capacity. (Vonk & Pitzen, 2017). While the empirical data are equivocal, if such a relationship does hold then differential levels or kinds of religiosity across communities may track variation in mindreading ability. The populations sampled here vary significantly in the proportion of participants with a religious affiliation, the diversity of religious affiliations among participants, and the importance with which one's religious affiliation is held. For example, the religious composition of Morocco at a national level is such that as of 2019, only 5% of the population describes themselves as non-religious or atheist. Of those who hold a religious affiliation, nearly all are Sunni Muslim, and 82 percent of the population describes themselves as either somewhat religious or religious (Arab Barometer, 2019). In stark contrast, nearly a third of the population in China identifies as atheist, only approximately 5% of the population belong to a religious organization, and more than 50% of the population identified as non-religious (Yao, 2007). While nearly three quarters of the population described themselves as having no religion

or practicing folk religion, the actual estimate of self-identified atheists is likely closer to a third of the overall population (Wenzel-Teuber, 2017; Yao, 2007). Unlike Morocco, there is a greater diversity of religious affiliation among the Chinese population writ large. While the largest percentages of the overall population identify as Buddhist (~16%) and Taoist (~ 8%), Christians (~2.5%) and Muslims (~0.5%) are represented as well (Wenzel-Teuber, 2017). If the religious composition of Morocco and China can be thought of as two ends of a single spectrum, the United States may lie somewhere in the middle of these two nations. Only 23% of the population does not adhere to a religion as of 2020. 70% of the United States population identifies as Christian, with 46% identifying specifically as Protestant and 22% as Catholic. A remaining 7% of the population adheres to a non-Christian religion, including Judaism, Islam, Hinduism, and Buddhism (PRRI, 2021). Nevertheless, nearly a third of the United States population identifies as not religious, underscoring the fact that affiliation and religiosity are distinct and dissociable parameters (WSJ/NORC, 2023). This all suggests a greater diversity of religious composition and an intermediate degree of irreligiosity when compared to Morocco and China. Taken together, such variation across samples increases the likelihood that any effect of religion may be observed. If religiosity is a meaningful predictor of mindreading ability, it is predicted that the greatest amount of mental state talk and strongest mindreading performance would be observed among Moroccan participants, followed by American participants, and lastly by Chinese participants.

Ethnicity

A large body of research has demonstrated that mindreading task response accuracy and self-reported confidence of participants is meaningfully predicted by whether or not the ethnic or racial identity of the participant matches that of the individual depicted in the task stimulus (Bradford et al., 2018; Moya & Henrich, 2016; Wu & Keysar, 2007). To the extent some aspect of the mindreading system operates on bodies of culturally inherited knowledge, and to

the extent ethnicity is a sufficiently informative index of one's possession of that knowledge, the ethnicity of a mindreading target may structure one's priors about their mental states. As such, communities that vary in ethnic composition may also vary in the frequency with which they must attribute mental states to those with whom they do not share such bodies of knowledge, thus influencing both the mindreading system itself and the language produced about the mental states of others. There is also a related methodological concern. Namely, participants recruited from more ethnically and racially homogenous societies may perform more poorly on the measures of mindreading I employed and may attribute fewer mental states when describing the video stimuli not because of a difference in competence, but because of a difference in how frequently they have had to make such attributions across such racial and ethnic lines. As such, this is a factor that may be meaningful to control for in analyzing the data generated in this dissertation. For these reasons, the populations sampled here represent interesting test cases with respect to their ethnic and racial composition.

In Morocco, people of Arab background constitute 44% of the population, with an additional 24% representing people of Arabized Berber background. Of the remaining population, 10% are Beidane and 1% belong to other racial or ethnic backgrounds (Laroui et al., 2024). Thus, approximately one half of the population belongs to one ethnic background and the remaining half belongs to another. In contrast, the racial and ethnic demography of China is nearly homogenous, with approximately 91% of the population belonging to the Han ethnic group. Of the remaining 9%, the only ethnic group that represents greater than one percent of the overall population is Zhuang (*China Statistical Yearbook*, 2022). While ethnicity is the predominant factor according to which social groups are differentiated in Morocco and China, the role of ethnicity is secondary to that of race in the United States. In the United States, the primary category of ethnic description is whether someone is or is not Hispanic or Latino. Here, approximately 19% of the population is Hispanic or Latino, while the remaining 81% are not.

Along racial lines, approximately 75% of the population is White, 14% of the population is Black or African American, 6% of the population is Asian, 1% is American Indian or Alaska Native, 1% is Native Hawaiian or Other Pacific Islander, and 3% of the population is Mixed Race or Multi-Racial (U.S. Census Bureau, 2020). Thus, while the United States may be more ethnically or racially homogenous than Morocco with respect to the proportion of the population represented by the majority group, it is less so than China. Additionally, the United States is composed of a greater number of ethnic or racial groups that constitute more than 1% of the population than China and Morocco. In essence, then, these three populations vary in their ethnic and racial homogeneity.

Language

Given my focus on the relationship between mental state talk and mindreading, the affordances of English, Mandarin, and Arabic for addressing questions in this space are manifold. It has long been suggested that languages encode concepts reflective of the values, beliefs, and perspectives of the cultures within which they emerge and are used. While some concepts might appear more regularly across languages than others, there are still likely others unique to particular cultures. Given the broad demographic, religious, political, and economic differences across these three societies, it stands to reason that the languages spoken in the United States, Morocco, and China may encode some of the values, beliefs, and perspectives associated with these cultural differences. Arabic, English, and Mandarin vary significantly in a number of dimensions, including linguistic typology, breadth of vocabulary, number of speakers, morphological complexity, and grammatical structure. This variation may entail differences in communicative practices and the concepts that are habitually encoded across these languages. If these differences pertain to the mind, then these three languages may collectively represent a meaningful set for testing differences in mental state talk.

Family size and composition

A number of studies conducted in WEIRD settings, a term developed by Henrich and collaborators to refer to “Western, Educated, Industrial, Rich, and Democratic” contexts (Henrich et al., 2010) have indicated that both birth order and the number of one’s siblings play a role in the development of mindreading (Lo & Mar, 2022; McAlister & Peterson, 2007; Perner et al., 1994). Crucially, however, these findings have not been replicated in societies that do not fit within this conceptual paradigm. Moreover, there is an emerging body of literature suggesting that the extent to which kinship structures are “intensive,” or are of a higher density, predicts the extent to which intentions figure in moral judgments (Schulz et al., 2019). If strictly true, then larger families might produce individuals with richer mindreading abilities. Therefore, participants recruited from Morocco, China, and the United States represent good candidates for comparison due to meaningful differences in average family structure. Despite a long cultural history which has placed an emphasis on the extended family, the introduction of the “one child” policy in China has resulted in changing family and household structures trending toward familial nucleation and fewer siblings within households (Chen, 1985). Though fertility policy has twice undergone changes in recent years permitting families to have multiple children, the one child policy stood as law for nearly 40 years (Su-Russell & Sanner, 2023). As such, a generation of Chinese citizens have grown up in households structured by its influence. Consequently, Chinese participants may overwhelmingly come from homes in which they were the only children. In Morocco, similar patterns of demographic transition have restructured average family and household composition in ways that parallel China, albeit more recently and without official policy (Berriane et al., 2021; Fargues, 2011). Whereas earlier patterns of residence emphasized the extended family and co-residence, families in Morocco are increasingly trending toward the nucleation seen in Western Europe and the Americas (Fargues, 2011). Crucially, these patterns have emerged more recently than in the United States and China.

Moreover, it has developed without official governmental intervention, like in China. As such, the total number of siblings and cousins with whom Moroccan participants grew up may exhibit greater variability. Finally, household structure in the United States has been in transition over the last 60, likely driving down the number of siblings with which prospective participants grew up. The nuclear family, composed of a married couple and their children, has been idealized as the "traditional" family structure within the United States since at least the 1950s, though it may extend back as early as 1880 (Ruggles, 1994). There is evidence to suggest that in the past 30 years, changes to family structure have leveled off such that 70 percent of children in the United States have two parents, though the percentage of American households composed of children living with both parents is only a quarter (Williams et al., 2012). As the demographic transition typical of post-industrial nations has impacted the United States, the birth rate is below replacement and the average American participant is likely to have one or no siblings (Smock & Schwartz, 2020). It warrants mention that these patterns are mapped only at the national level and there exists significant variation according to socioeconomic, regional, and individual factors. Nevertheless, the broad variation across these samples may permit observation of variation in mindreading and mental state talk among participants drawn from the United States, China, and Morocco. Specifically, if the number of siblings one has and the number of members in one's household both positively predict mindreading ability and production of mental state language, such effects may be observed most strongly in Moroccan participants, followed by American participants, and lastly by Chinese participants.

Materials

Videos

A total of nine short (approximately one minute in length), silent video stimuli depicting naturalistic interactions between two or more individuals across a variety of contexts and situations were created in order to collect samples of elicited speech. Specifically, participants

described the vignettes they had seen to the experimenter as if they were telling a friend about something they had actually seen. Though these descriptions may imperfectly represent unprompted, naturalistic retellings of events, they provide a simulacrum of such speech acts and, more importantly, control for speech content across samples, thus affording more standardized observations of LR3PMS across languages.

Several considerations went into designing the video stimuli used to elicit mental state talk. First, they were designed to depict scenarios understandable and interpretable to people from as broad a range of cultural backgrounds as possible. Ethnically and racially ambiguous actors were cast to maximize participants' credulity that the stimuli depicted people with whom they might actually interact. Where access to actors who fit this description was limited, actors from a variety of racial and ethnic backgrounds were cast. This same logic motivated the removal of audio from the video stimuli. By designing stimuli interpretable without hearing the actors' dialogue, they were unlinked from any one particular linguistic environment and could thus be used across a broader sample of populations. The depiction of highly culturally-specific technologies, tools, clothing, environments, and artifacts was limited by employing only those with a long history of use across the world (such as axes and soccer balls), instructing actors to wear minimally branded or decorated clothing (i.e., unmarked t-shirts and jeans), and setting the stimuli in predominantly natural settings with as few constructed features as possible. The stimuli were constructed to depict simple social interactions where inferences about the motivations, desires, goals, and other internal states of the actors might be useful or necessary for understanding their behavior, thus drawing upon the mindreading capacity. As such, the frequency of lexical references to third-party mental states (LR3PMS) in participants' descriptions may constitute a meaningful measure of its deployment in speech. Importantly, participants were not prompted to use MS language, or to make their descriptions mentalistic (see prompts below). Finally, the stimuli were designed to depict categories of social interaction

that have been the focus of evolutionary research on human behavior, such as mate competition, status competition, and cooperation. Given that human psychology may have undergone selection to preferentially allocate attention to and interest in social interactions whose outcomes influence fitness, if such selection has operated uniformly across human populations, then attention to and interest in video stimuli of those interactions should be uniform across populations as well. Drawing on literature centering the role of mindreading in human fitness across contexts like cooperation, deception, and resource acquisition and protection (Barrett et al., 2010; Cheney et al., 1986; Emery & Clayton, 2001; Henrich & Gil-White, 2001; Lyons & Santos, 2006; Paal & Bereczkei, 2007), a sketch of eight fitness domains thought to interface with mindreading was drawn. These domains were cooperation, dangerous animals, dominance, infidelity, mate guarding, norm violation, prestige, and sickness. Narratives for each were written and thus served as the basis for each stimulus.

A ninth script was written depicting a situation modeled on the False Belief test (Baron-Cohen et al., 1985; Wimmer & Perner, 1983). While a departure from the design logic of the other eight stimuli, the False Belief test is a tool which has been widely used to measure individuals' abilities to track others' mental states (Dennett, 1978). Therefore, this stimulus served as a baseline against which to compare the other videos. The narrative arc of each video stimulus was written to minimize reliance on character dialogue and to ensure that the action in the story hinged on understanding the characters' mental states. All audio was removed in post-production, though tones were added to indicate the start and end of the video. Together, these design constraints aimed to ensure ready interpretation of the video stimuli across linguistic and cultural groups. In the following sections, summaries of each video are provided.

Cooperation. Two female actors (Actors 1 and 2) are each trying to move large and heavy objects from the ground onto tables. The first woman is trying to lift a large box and the other is trying to lift a large pot. They are facing away from each other and working at different

locations. Each woman struggles to lift her respective object. After a short time, Actor 1 gives up and notices Actor 2. Actor 1 approaches Actor 2 and gets her attention. Actor 1 helps Actor 2 lift the box onto the table and then points to her pot while smiling. Actor 2 starts to unload her box and ignores Actor 1. Actor 1 points again at her pot and Actor 2 shoos her away while continuing to unload her box. Actor 1's face looks angry and she walks away, exasperated. She returns to her pot and, with great effort, lifts it onto the table (**Figure 1A**). A wide-ranging literature has emphasized the importance of cooperation in human beings' evolutionary success (Boyd & Richerson, 1992; E. Fehr et al., 2002; Richerson et al., 2016). Moreover, there is a rich literature on the fundamental intersection of mindreading and cooperation, with authors emphasizing the ability to take others' perspectives and establish joint goals (Barrett et al., 2010; Caballero et al., 2013; Paal & Bereczkei, 2007; Sally & Hill, 2006). These dynamics were likely true of the last common ancestor of all extant human populations and as such, the kinds of mindreading one needs to do in such circumstances ought not to vary across cultures.

Dangerous animal. Two female actors are walking outside together. They stop beneath a tree branch and begin speaking to each other. As they are speaking, Actor 1 notices a snake in the tree branch directly above Actor 2's head. She stops talking and backs away from Actor 2 while looking intently above her. Actor 2 looks puzzled and begins to look around. Finally, she notices the snake above her and jumps back, startled. Actor 1 now looks afraid and calls Actor 2 toward her. Actor 2 runs over to Actor 1 and they look for a stick. Using the stick, they knock the snake out of the tree and kill it. The two actors then run off camera, presumably to look for help or warn others (**Figure 1B**). The threat posed by poisonous, venomous, and predatory animals represents a selection pressure whose existence long antedates the evolution of anatomically modern human beings. As such, circumstances depicting interactions with such threats ought to be readily interpreted by viewers independent of cultural background. Good theoretical accounts

have suggested that mindreading is an essential component of predator-prey relations (Barrett, 2005). Thus, viewers may be primed to discuss the actors' attitudes and representations.

Dominance. A male actor is shown squatting next to a pile of cut wood. The man appears tired and wipes sweat from his forehead. A second, larger male actor then slowly saunters toward Actor 1 and his pile of wood. He smirks and gestures menacingly at Actor 1, pounding his chest as he walks past him and attempting to steal his wood. Actor 1 places his ax over the pile to prevent Actor 2 from stealing the wood, but Actor 2 grabs the ax and the two men struggle over it. Actor 2 wrests the ax from Actor 1's hands and gathers the wood and begins to walk away. Actor 1 pleads with Actor 2 by grabbing his arm. In response, Actor 2 turns around and threatens Actor 1 with the ax. Actor 1 backs away and continues to plead with Actor 2. Actor 2, without looking back, throws a single piece of wood on the ground in the direction of Actor 1. Actor 1 looks despondent (**Figure 1C**). Much like the previous scenario, dominance relations are selection pressures whose influence has likely characterized human evolution far beyond anatomically modern *Homo sapiens*. These kinds of interactions have fitness impacts across a wide range of social, group-living species (Cheney et al., 1986; Cheng et al., 2013; Henrich & Gil-White, 2001) and as such, the structure of such interactions ought to be readily understood across cultural contexts. Moreover, the resource at hand is one whose function is fairly universal and whose production requires much the same work across cultural contexts.

False belief. A male actor approaches a clearing in the reeds near a small river. He feels the water with his hands before removing his shirt to bathe. He realizes he has forgotten his soap and walks out of frame. To screen left, there is a small white tub full of clothes on the ground. A woman approaches the tub and begins preparing to wash its contents. As she is removing clothes from the tub, she notices the man's shirt on the tree. She takes it and adds it to the tub, but realizes she has forgotten some cleaning supplies. She leaves with the tub. The man then enters stage right and looks confused by his shirt's disappearance. He looks around

and the woman enters from stage left. They briefly chat and she returns his shirt to him, realizing it was a misunderstanding. This circumstance was included as a check to compare against the other scenarios (**Figure 1D**). While there is no evidence to suggest that scenarios structured by their participants' false belief have operated as important selection pressures in human evolutionary history, the False Belief task is a well-validated measure of mindreading in the literature (Wellman et al., 2001; Wimmer & Perner, 1983). Therefore, it is reasonable to expect that this circumstance would elicit mental state language.

Infidelity. A female actor (Actor 1) and a male actor (Actor 2) are seated on a sofa together with a door visible in the frame. Actor 1 and Actor 2 are sitting close together and holding hands. Both appear to be very happy. As they are sitting, the door opens and Actor 2 jumps up from the sofa. A second female actor (Actor 3) has entered the room with bags in her hands. Actor 3 drops the bags and begins to scream at both Actor 1 and Actor 2. Actor 2 looks guilty and surprised, while Actor 1 looks increasingly uncomfortable. She gets up off the sofa and walks toward the door while Actor 3 berates her. Finally, once she has left, Actor 2 and Actor 3 begin to get into a shouting match (**Figure 1E**). In any species with high degrees of parental investment, infidelity represents a threat to ones' fitness. Because the time and energy invested in the relationship is zero-sum, any investment in extra-pair interactions comes at a cost to the primary relationship (Schaffer, 1974). Thus, emotional reactions like anger to such costs are thought to be fairly universal across cultures (Buss et al., 1992; Daly et al., 1982). Therefore, the structure of this problem ought to be readily interpreted across field sites.

Mate guarding. A female and a male actor (Actor 1 and Actor 2, respectively) are seated on the ground in a grassy area. They are smiling and laughing. Another male actor (Actor 3) can be seen walking at a distance. Actor 1 smiles, gets up, and walks toward Actor 3. She hugs him and talks animatedly with him, touching his arm and being flirtatious. Actor 2 is left sitting alone. He appears increasingly confused and angry. Eventually, Actor 2 gets up and walks over to

Actors 1 and 3. He attempts to introduce himself but is largely ignored by Actor 3. Actor 2 conspicuously places his arm around Actor 1's neck. Despite this signal of his discomfort, Actors 1 and 3 continue to talk. Actor 2 eventually guides Actor 1 away from Actor 3, who continue talking as they depart (**Figure 1F**). Like infidelity, an abundance of research on human sexual behavior and its associated psychological mechanisms has suggested the adaptive value of emotions like jealousy in these contexts (Buss et al., 1992). Some evidence suggests sexual jealousy may be universal (Buss et al., 1999; Buunk et al., 1996). As such, this scenario is likely readily understood by participants from a variety of cultural backgrounds.

Norm violation. A group of three actors (two female actors and one male actor) are standing on a small, elevated platform together. A series of three actors (two male actors and one female actor) enter and give gendered gifts to each of the three actors standing on the platform. Each female actor receives a rose from the actors who walk past, while the male actor receives a bottle of beer. The first actor to walk past is a man, and after giving his gifts he stands off to stage left. The next actor is a woman, and she lines up with him after having given her gifts. The third is a younger man, and he gives the wrong gifts. He appears to realize he is doing it incorrectly, and smirks when he lines up with the other two, looking at both of them to gauge their reactions (**Figure 1G**). Many human behaviors, practices, and institutions are structured by arbitrary rules selected from a broader range of possible rule sets (E. Fehr et al., 2002). Adherence to these rules is often moralized and failure to do so is sanctioned, even when the material consequences of such violations are minimal or non-existent. While the specific rules that exist across societies vary, the structural features of such scenarios may be a universal feature of human groups. As such, this scenario may still be interpreted through the lens of a norm violation by participants from many cultural backgrounds.

Prestige. Two male actors are shown standing in a field. Actor 1 is standing with a soccer ball and authoritatively pantomiming how to juggle a soccer ball while Actor 2 looks on

intently. Actor 1 begins to demonstrate juggling but is not able to do so. He looks embarrassed briefly but insists Actor 2 continue to watch him. As Actor 1 continues to try and teach Actor 2 how to juggle, a third actor enters the scene in the background. Actor 3 is far more skilled, juggling the ball with ease. Actor 2 notices Actor 3, but Actor 1 insists Actor 2 continue to watch him. After a short time, Actor 2 tells Actor 1 he is going to talk to Actor 3. He departs and introduces himself. Actor 1 is left standing by himself. He waves his arms to get Actor 2's attention, but he is focused on Actor 3 (**Figure 1H**). As human beings came to rely increasingly on cultural innovations to exploit the niches they occupied, expertise in those cultural practices became an essential fitness currency (Cheng et al., 2013; Henrich & Gil-White, 2001). In a given domain, those who attend to the behavior of successful individuals are able to copy their methods, thereby increasing their own success. While soccer may not bear on one's fitness in the same way as learning to extract some resource, it is a domain in which individuals' skills vary. The global popularity of soccer suggests that this scenario should be readily interpreted.

Sickness. A female and male actor (Actors 1 and 2) are shown standing together talking on one side of the frame. On the other side, a female actor (Actor 3) is leaning against a tree. She holds her head and hunches over, appearing ill. Actors 1 and 2 notice her but continue talking. Eventually, Actor 3 vomits and collapses. Actor 1 runs over to her assistance and, upon arrival, gestures for Actor 2 to join her. He hesitates, but eventually comes over. They help Actor 3 stand up. Actor 1 then escorts Actor 3 out of the frame. Actor 2 appears disgusted and wipes his hands on a tree before following them out frame (**Figure 1I**). Like infidelity, mate guarding, and dangerous animals, the fitness consequences of illness apply to many species. However, humans are unique in the extent of care provided to the infirm (Carter, 2014). Also, the affective experience of disgust may have evolved to structure behavior in a way that limits pathogen exposure (Tybur et al., 2013). Evidence suggests these behaviors are widespread across human societies and should be interpretable to participants from many cultural backgrounds.

Attention check and mindreading questions

In addition to providing descriptions of all nine video stimuli, participants were asked to answer a series of three questions about the last 4 or 5 of the videos, depending upon the block into which they had been placed. Of these three questions, the first was a simple attention check question and the remaining two asked explicitly about the mental states of the agents depicted in the video. These questions were included for two reasons. Attention check questions were included as a low-resolution means by which to exclude data. These questions were easy for participants to answer if they attended to the video, and incorrect answers were thus taken to suggest the participant had not paid attention. Mindreading questions were included to more explicitly target mental-state speech and to assess participant mindreading ability. If participants did not produce mentalistic descriptions of the videos, they might nevertheless do so when asked directly about a character's motivations, beliefs, and desires. These questions allowed me to assess that possibility. A list of these questions, their tentative answers, and data suggesting their level of difficulty can be found in Appendix A.

Reading the Mind in the Eyes

Participants were also asked to complete the "Reading the Mind in the Eyes" test (Baron-Cohen et al., 1997, 2001), a widely used measure of emotion recognition, a mindreading capacity which has shown to vary among neurotypical adults. This test is composed of thirty-six images of eyes derived from print advertisements, each of which is surrounded by four words. These words are candidate descriptors of the affective or mental state of the eyes depicted. During its initial development, a panel of four researchers discussed each of forty images to come up with a single term best describing each image. Three foil items were also proposed for each image. These images were then passed to a panel of eight raters tasked with choosing the "correct" word for each of the forty images. If the raters failed to unanimously select the "correct" item as determined by the initial panel of researchers, the image was returned to the

researchers for revision and subsequent re-evaluation by the panel of eight raters. This was done until the panel of eight raters had unanimously arrived at the correct answer as chosen by the researchers for all forty images. During pilot testing with participants in surrounding communities, Baron-Cohen et al. (2001) dropped four of the images from further use as participant responses failed to meet the item inclusion criteria that the modal response must represent greater than fifty percent of the total responses and the second-most common response must not be greater than twenty-five percent of all responses.

The Reading the Mind in the Eyes Test thus instantiates several qualities desirable in a tool to measure individual differences in mindreading across cultures. First, it was designed for the express purpose of examining individual differences in the mindreading abilities of neurotypical adults and has sufficient resolution to detect these differences between individuals. Second, the process by which the test was created is amenable to cross-cultural tuning. While a growing literature suggests that individuals tend to be less accurate in mindreading tasks that depict out-group as opposed to in-group members, these data do not necessarily pose a problem for the Reading the Mind in the Eyes Test. Whether or not these data pose a problem depends on the nature of the error. If errors in inter-ethnic mindreading are systematic such that all members of a community X make the *same* erroneous mental-state attribution to a member of community Y, then so long as the terms surrounding each image in the Reading the Mind in the Eyes Test have been developed by a panel of individuals from community X according to the procedure specified by Baron-Cohen et al. (2001) it is irrelevant whether they reflect the “actual” mental state. By this same logic, it may even be irrelevant if the terms surrounding each image have been developed by a panel of individuals from community X. To the extent that all members of a given community have a similar interpretative framework through which to understand emotional expressions, and to the extent they select the best candidate descriptor of the four words already provided, participants ought to rank them similarly with respect to the

strength of their fit to the image. The criterion for success has always been the extent to which one's interpretation of an emotional expression matches the consensus interpretation, *not* how accurate that interpretation is. Having extracted the stimuli from advertisements, it is not the case that the mental states of the individuals depicted therein could even have been confirmed independently. It has always been a task of interpretation, though one that has emphasized agreement in interpretation. The use of the same set of stimuli may, in fact, be a virtue of this study, as participants across contexts have viewed the same images and thus controls for their particular effect.

Third, the Reading the Mind in the Eyes Test is one of only a few tools for measuring neurotypical adults' mindreading ability (cf. Turner & Felisberti, 2017) and of them, it is among the most well-documented and empirically validated. It is for these reasons that I employed this metric as an indirect measure of individual mindreading ability. Using this tool, it is possible to assess whether or not lexically encoded references to others' mental states are predicted by some subcomponent of the mindreading capacity – namely, emotion recognition. Moreover, it is possible to examine whether there exist cross-cultural differences in this relationship. More specifically, the mindreading abilities assessed by the Reading the Mind in the Eyes test represent those that use relatively impoverished data about an interlocutor's face to impute a representation of their affective state. However, it also implicitly tests the efficacy with which that representation can be used by the language system to pair it with appropriate lexical-semantic representations, against which the candidate descriptors are compared. It is likely the case that at least some portions of these abilities are different than those used to represent the mental states of agents in a false belief task or those that initiate a fear response after perceiving a pair of eyes trained on oneself. Whatever differences or similarities are found across cultures with respect to performance on the Reading the Mind in the Eyes Task may not generalize to other mindreading skills.

Software

Given the nature of the data this study aimed to collect, software was not strictly necessary. In principle, all that was required was a device to record participant audio and a means by which to display video stimuli to participants. However, such an approach increased the probability of experimenter error and introduced substantial processing demands following data collection. To circumvent these issues, two distinct sets of software were used to collect data according to whether participants were interviewed virtually or in-person. These tools were implemented to accommodate restrictions placed on in-person research conduct in response to the global COVID-19 pandemic. Thus, the software tools used by collaborators to implement the study in-person (prior to COVID-19) consisted of Open Data Kit (ODK) Collect, an open-source Android application for conducting surveys and interviews when disconnected from wireless networks (as is commonly the case in field research settings), and ONA.io, a web-based platform capable of interfacing with ODK and serving as a remote server onto which completed surveys could be stored upon re-establishing wireless network or internet access. The tools used to conduct the survey virtually (after COVID-19) were the experimenter's choice of videoconferencing software and a custom program written in Python that could flexibly interface with any such platform. This program facilitated and standardized data collection while also minimizing and streamlining subsequent data processing. In both cases, the software was used to collect participant demographic data, play the videos in a pseudo-randomized order to participants (a design decision borne out of limitations on true randomization inherent to the function of ODK Collect), record audio of elicited narrative descriptions, and mark responses to attention check and mindreading questions. Across both implementations, the structure of the study design was unchanged and differences reflected accommodations made for virtual study conduct.

Design

The current study is a mixed design with 'video' as a within-subject factor and 'culture' as a between-subject factor. Participants viewed a series of nine videos, provided descriptions of the videos, and answered questions about the videos according to one of eight different pseudo-randomized orders. Because the questions had not undergone prior validation, as well as concerns about their influence on guiding attention during viewing, two blocks of videos were presented to participants. In the first block, participants only described the videos they saw. In the second block, participants described the videos they saw and answered three questions about the videos. The first block included the False Belief video (see Materials section for more detail) for one half of the participants, while it was in the second block for the other half[1].Table 1 details the eight conditions and pseudorandomized orders into which participants were placed.

Procedure

In both virtual and in-person interviews, the experimenter and participant were seated facing each other. Participants provided verbal consent to record the interview to be made for later review and analysis. In both virtual and in-person interviews, two recordings were made simultaneously – one using the data collection software and one using some other tool to record digital media. The production of two recordings was implemented as a safeguard against software or experimenter error. Participants were randomly assigned to one of eight pseudorandomized conditions determining the order of video presentation. Video presentation order was pseudorandomized, and not truly randomized, due to design limitations in ODK Collect. After pseudorandomized condition assignment, a range of demographic data were

[1] Given prior theoretical considerations regarding the False Belief task as a standardized narrative against which mindreading is measured, and the aforementioned concerns about questions guiding attention during viewing, the study design is organized so as to allow a subsample to provide narrative descriptions of the False Belief video without having primed the participants to attend to specific features of the False Belief video. This allowed comparisons across orders to see if the questions had, in fact, influenced the nature of participants' descriptions.

collected from each participant (See Appendix B for survey example) and logged in the survey software. The experimenter then described the study procedure and provided instructions to participants about how they should think about and frame their descriptions of the videos they were to see. Specifically, participants were encouraged to describe what they had seen “as if they were telling a close friend about something they had actually encountered” in order to approximate naturalistic everyday speech.

Participants then viewed the first video stimulus. When the tone indicating the end of the video sounded, the experimenter asked the participant whether they had encountered any technical difficulties viewing the video and if they would like to view it again for any reason. These questions also provided an opportunity to determine whether there were systematic differences in understanding between those who requested second viewings and those who did not. If participants encountered technical difficulties, they were encouraged to watch the video again. If participants wished to re-watch the video, they were permitted to do so as many times as they liked. Of the 177 total participants, only 9 participants rewatched any videos at all. Of these 9, five rewatched only a single video. A single participant rewatched two of the videos, two participants rewatched three of the videos, and a single participant watched every video twice. In short, three participants accounted for nearly 70% of the rewatches. The distribution of rewatches across videos appeared random, suggesting no systematic issues of interpretability.

After viewing a video, participants were instructed to wait for the experimenter to indicate that the software was recording before beginning their description of the video stimulus. Participants then described the video in as much or little detail as they wished and were reminded to describe the event as if they were telling a close friend about something they had actually seen. Experimenters were instructed not to give any positive or negative feedback on participants’ descriptions, and not to suggest the participant slow down, hurry up, or stop. This procedure was repeated for all remaining videos in the first block of the pseudorandomized

condition. During the second block, the procedure was identical except for the inclusion of three follow-up questions about each video (See Appendix A for questions).

Upon completion of the elicited video descriptions, participants completed the Reading the Mind in the Eyes task. Participants were sequentially presented with 37 images of eyes, each of which was surrounded by four candidate descriptors. Participants were asked to pick one of the four words to describe the emotion depicted in the image they saw. The first image provided participants an opportunity to familiarize themselves with the procedure. During the presentation of this first image, participants were informed that the experimenter had a list of definitions for all the words they would encounter and that they should not hesitate to ask for the meaning of a word if they were unfamiliar with it. As a forced choice task, participants could not skip items.

The audio of the elicited narrative descriptions and question responses were transcribed using the Google Cloud Speech-to-Text API and a series of Python scripts written to automate the process. Afterwards, two research assistants per language reviewed these automated transcripts in order to ensure their accuracy. Where self-interruptions or incomplete terms occurred (i.e., "I s-, I saw"), only those that were complete were included (I, I saw). Non-linguistic utterances like laughter and sighs, as well as other components of speech like pauses were not transcribed in the present studies but may be transcribed and analyzed in the future. See Appendix C for a detailed account of how data were processed from their raw form into a format appropriate for coding by research assistants.

Coding

For each language examined, a Python script was run which catalogued all unique lexical items produced across all of the transcripts for each of the three target languages. This script cataloged both the types (unique items) and tokens (counts of each unique item) of the lexical items in each language-specific corpus of transcripts. The resulting spreadsheet of this

catalogue, referred to as the Dictionary File, and a set of instructions for identifying lexical references to third-party mental states (LR3PMS) were provided to no fewer than two coders per language. For each lexical item type, coders were tasked with coding each item in the Dictionary file according to whether or not it could reasonably be glossed as potentially referring to a mental state. Coders made this evaluation based upon the set of instructions provided to them, which defined the targets of coding according to the specific criteria of the particular study to which their work corresponded. After each coder for a given language had completed this step, their Dictionary Files were fed into another Python script that used the list of coded lexical item types to label all of the corresponding tokens across all transcripts in the language's corpus of speech samples. This document, known as the Raw Data File, thus featured all tokens of potential mental state terms in their original speech context. Because this process was undertaken for each coder separately, two distinct Dictionary and Raw Data files were made for each language sampled.

Raw Data file review. The Raw Data file was generated according to this procedure for a variety of reasons, including to reduce human error, increase speed, and ensure all instances were actually captured. That is, this procedure sought to minimize the likelihood of false negatives, a problem that had presented itself in early piloting of coding procedures. However, it also increased the likelihood of false positives**.** Given the column in the Raw Data file containing candidate LR3PMS had been populated automatically, it is all but guaranteed that a subset of the coded items included tokens that were erroneous, inaccurate, or mismatched to the criteria for an LR3PMS. Without any human review of the coded tokens, they could erroneously include false positives instead of just genuine instances of LR3PMS.

An example of one way in which this might occur can be seen in the failure of English to differentiate verbs morphologically when conjugating for 1st person singular and 3rd person plural in the present tense. Because the script used to populate the Raw Data file from the dictionary

would search blindly for instances of “think” (presuming it had been coded by a coder in their Dictionary file), “think” as used in the sentence “I think that this video was strange” and “think” as used in the sentence “They think the man was behaving badly” would both be coded. Another way in which this might occur is when a mental state term is homonymous with another that does not refer to mental states. For example, the word alert could refer both to the mental state of clarity and energy as well as a warning signal, the former of which could represent a mental state and the latter of which does not. In order to ensure the items coded in the Raw Data file were correct, coders were tasked with reviewing each coded token in context. The goal of this review was to pare down the set of coded items to only those that constituted LR3PMS. One complicating factor faced by coders in their review of the Raw Data file is that of lexical references to first-party mental states that occur in the context of playacting or taking the perspective of third parties depicted in the video stimuli. As such cases represent first-person mental state references with respect to the grammar or morphology of the target language, they might seem at first glance to be candidates for removal. However, it is unlikely that a participant describing a video stimulus would be able to embody the perspective of a character depicted therein and ascribe to themselves, in the role of the character, a mental state without first having attributed it to that character while viewing the video. As such, it is possible such speech acts draw on the same cognitive mechanisms required to make third-party mental state references outside of playacting or quotative contexts (Goldstein & Winner, 2011; Taylor & Carlson, 1997). Even though the inflection of the coded item may not have indicated the grammatical third person, the act of taking the character’s perspective and producing a mental-state term in their voice requires attribution of the corresponding mental state. For example, a participant describing the Mate Guarding video stimulus might say something to the effect of, “…and then the guy with the jacket came over and was like, ‘I don’t believe what I’m seeing right now – I thought she agreed not to talk to her friends like this. I know she remembers that discussion!’”.

Here, four instances of LR3PMS would be coded – believe, thought, know, and remembers. If, however, a participant describing the Mate Guarding stimulus said something like, “and then the guy with the jacket came over and I believe he was upset at the girl. I thought he was maybe jealous of the other guy, but who remembers exactly what happened before he got up”. Here, the same four tokens would not be coded as L3RPMS because they all refer to the mental states of the participant themself. Though some of the tokens in the first and the second sentence may be of the same grammatical class, their usage in context differentiates them with respect to the kind of mindreading involved. In effect, because it is not the participant’s mental state referred to in quotative speech, but that of the character, such cases were taken to fit the criterion of a LR3PMS.

Consequently, coders could not rely on the grammatical or morphological cues of the coded items alone to determine their eligibility for inclusion in further processing and analysis of the Raw Data file. To that end, coders were tasked with reviewing each positively coded item by hand, reading as much of the preceding and proceeding text surrounding the coded word token as was required to determine whether it constituted a genuine instance of an LR3PMS, be it in the context of direct or quotative speech. If a word token was deemed by the coder to fit the criteria for an LR3PMS, the code was left unaltered. If, however, the word token was deemed to constitute a false positive, the code for word token was altered to remove it from the set of LR3PMS. Coders proceeded according to these steps for all coded word tokens in their respective Raw Data spreadsheets until they arrived at the end of the document. Once each coder had completed their review of the coded word tokens in their respective Raw Data files, the documents were shared with the lead experimenter in order to run inter-rater reliability analyses and evaluate whether additional rounds of data processing were required.The novel methodology presented here represents a crucial development toward addressing some of the outstanding questions pertaining to the relationship between mindreading and language. In the

following chapters, the data generated by this approach were coded, analyzed, and presented with the goal of determining, for the first time, if speakers of different languages differed meaningfully in the frequency with which they produced LR3PMS.

Chapter 3: Examining Cross-Linguistic Variation and Uniformity in the Production of Belief-Like Mental State Verbs

Introduction

In this chapter, I assess whether lexical references to third-party mental states (LR3PMS) varied across a standardized corpus of narrative descriptions of video stimuli collected from participants in China, Morocco, and the United States. Participants from each field site were first-language speakers of Mandarin Chinese, Moroccan Arabic, and American English, respectively, and their speech samples were produced in these three target languages. Lexical references to third-party mental states (LR3PMS) were here defined as all instances of a predetermined inventory of mental state verbs derived from Wellman and Estes (1987) that were used to refer to the minds of characters depicted in the video stimuli. LR3PMS included both verbs conjugated for the third person used to describe the mind of a character as well as verbs conjugated for the first person used in the course of quotative speech – that is, instances in which the participant play-acted the speech of a character depicted in the video stimuli and used a first-person form of the mental state verbs in the inventory to refer to the mind of a character. Though these two forms of LR3PMS map onto distinct grammatical cases, they are treated here as members of a cohesive semantic class characterized by the imputation and subsequent attribution of unobservable mental states to agents distinct from oneself. Under this view, quotative LR3PMS and references to one's own mental states may be grammatically identical, but they differ with respect to the observability of the referent mental state to the speaker. Thus, quotative speech may be understood as similar to third-person LR3PMS by virtue of a shared reliance on the mindreading system to impute and produce language about the mental states of others.

The importance of an answer to the question of whether LR3PMS vary across cultural and linguistic contexts cannot be understated. Research across the fields of anthropology,

psychology, cognitive science, and linguistics have all variously contended with the relationships between language, culture, and mindreading (Bradford et al., 2018; Bretherton & Beeghly, 1982; Brooks & Meltzoff, 2005; De Rosnay et al., 2014; Dixson et al., 2017; Hawkins & Goodman, 2016; Hughes et al., 2014, 2018; Lecce et al., 2021; K. Milligan et al., 2007; Ruffman et al., 2002). To date, there is an abundance of work that has explored whether language about the mind varies cross-linguistically (Cheung et al., 2009; Devine & Hughes, 2019; Durrleman et al., 2019; Goddard, 2010; Heyes, 2018; Hoffman et al., 1986; Jackson et al., 2019; Kockelman, 2006; Levinson et al., 1987; K. Milligan et al., 2007; Pinto et al., 2017; Robbins & Rumsey, 2008; Ruffman et al., 2002; Salmond, n.d.; Schieffelin, 2008; Schwanenflugel et al., 1994; Sperber & Wilson, 2002; Stivers et al., 2011), whether social practices about the mind, such as moral judgments about blameworthiness, vary cross-culturally (Barrett et al., 2016; Heyes, 2018; Hughes et al., 2018; Lillard, 1998; Matsumoto, 1989; Schulz et al., 2019), and whether mindreading varies across human populations (Bradford et al., 2018; Gendron et al., 2014; Kuntoro et al., 2013; Perez-Zapata et al., 2016; Slaughter & Perez-Zapata, 2014). In addition, there is also rich theoretical and empirical work that explores whether causal relations can be said to exist between them (Boroditsky, 2011; Gumperz & Levinson, 1991; Haspelmath, 2010; Huettig et al., 2010; Phillips & Boroditsky, 2003; Tajima & Duffield, 2012; Wu & Keysar, 2007).

Given the complexity of the phenomena at hand, there are many ways in which the causal relationships between and the variation (or lack thereof) present within language, culture, and mindreading might manifest. Much of the extant research which bears on these questions has so far been conducted in piecemeal fashion with respect to the triadic relationship among these phenomena. Though variation in a single construct, like language (Goddard, 2010; Jackson et al., 2019; Levinson et al., 1987; Stivers et al., 2009), or the existence of a relationship between two constructs, like culture and mindreading (Adams Jr et al., 2010; Slaughter & Perez-Zapata, 2014; Wu & Keysar, 2007), has been documented, these findings

have not been the result of studies designed to assess causal relationships between and variation within the other relevant phenomena, nor have they been interpreted through such a lens. This state of affairs is likely a consequence of the absence of a complete and systematic inventory of models which aim to describe the variation within and causal relationships between language, culture, and mindreading. Such accounting of the possible world of explanatory causal models provides a framework within which to situate extant findings, guide the design of future research, shape the methods by which new data is collected, and weigh the plausibility of competing models according to their concordance with the data.

It is important to be clear that the current research, as in the case of the extant findings in the literature, cannot speak directly to nor provide positive evidence in favor of any such models of causal relations between language, culture, and mindreading. The reasons for this are manifold, but among the most important are the facts that language and culture do not vary independently of each other in the sample of participants recruited for this work, nor was any intervention performed to manipulate these variables. Consequently, any results described herein are simply correlations and as such, do not necessarily entail causation. Nevertheless, the current research may provide evidence against some of these accounts, narrowing down the set of plausible models and moving the field's collective understanding of these relations forward in a productive manner. Though correlation does not entail causation, arguments from the literature on causal discovery as well as some interventionist accounts of causality suggest that causation does entail mutual information, or a degree of mutual dependence between two variables. This can be understood as correlation broadly defined, at least somewhere in the causal chain between two relevant variables (Woodward, 2005). As such, if causation entails correlation, and correlation is not observed anywhere in the causal chain between the two variables, the absence of causation is logically entailed. It is according to this logic that the

present study may winnow down the plausible models of the relationships between and variation among language, culture, and mindreading.

Of all the causal explanatory models contained within the possible world of relations between language, culture, and mindreading, the most complex of them is one in which bidirectional causal arrows point from any one of the nodes represented by language, culture, and mindreading to all other nodes in the graph representing the model. A presumption of this maximally complex model, as well as many of the other possible models entailed by such directed acyclic graphs of their causal relationships, is that there exists variation in some or all of the parameters each node in the model represents. Variation in language about the mind can only cause variation in these other constructs to the extent that variation in language about the mind actually exists. The same can be said for cultural practices about the mind and for the mindreading capacity itself. In effect, the extent to which scholars ought even to concern themselves with such models depends upon the existence of variation within each of these constructs in the first place. Though there exists a small body of literature which purports to document variation in LR3PMS, the findings reported therein are vulnerable to a number of criticisms limiting the credulity with which they should be taken. Ruffman et al., (2002) have shown that parents in English-speaking households across the United States, Canada, and Australia vary in the frequency with which they produce LR3PMS, independent of overall speaking time, and that this variation predicted the age at which their children first passed the False Belief Test (Baron-Cohen et al., 1997). Similarly, ethnographic research findings in anthropology have suggested that there exist societies which explicitly prohibit talk about the minds of others (Duranti, 2008; Robbins & Rumsey, 2008; Schieffelin, 2008). While these data are a suggestive first step toward building theories that account for causal interactions between and variation within language, culture, and mindreading, the predominant critique of these studies is that they are neither clear about nor replicable in their methodology. The ethnographic

research pertaining to this question has been produced almost entirely by scholars who were not raised as members of the communities within which they work. Despite the unequivocal richness of their ethnographic insight and experience, the etic perspective such scholars bring to bear, coupled with the qualitative nature of their observations, limits the strength of their claims. Without systematic and replicable methods, the findings cannot be independently corroborated. More importantly, researchers operating from an etic perspective may generate data replete with false negatives by missing genuine instances of LR3PMS glossed as such by insiders but overlooked or misunderstood by outsiders.

The current research involves the analysis of a dataset collected according to the methods detailed in “Chapter 2 – General Methods” and focuses on devising and deploying a first-pass coding scheme by which to meaningfully compare talk about the mind of others across languages. In so doing, this chapter can contribute to the literature by validating or disconfirming the presence of cross-linguistic variation in LR3PMS. To do so, it is important to address a few unsettled questions. Namely, what is a mental state? And what words are used to refer to them?

Though the literature is replete with candidate answers to both of these questions, the early mindreading literature provides an effective starting point for the development of a coding scheme according to which corpora of speech samples may be processed. This approach is motivated by early debates in the field of cognitive science which remain relevant to the questions this dissertation seeks to answer. Specifically, early cognitive scientists interested in mindreading tended to treat mental states as representations, and representations as equivalent to or synonymous with propositional attitudes (Bretherton & Beeghly, 1982; Dennett, 1978; Fodor, 1992; Gopnik & Astington, 1988; Gopnik & Wellman, 1992; Leslie, 1987; Leslie & Happé, 1989; Perner, 1988; Wimmer & Perner, 1983). Among philosophers of mind, propositional attitudes can be understood as causally efficacious, content-bearing internal states (Nelson, 2023). That is, a propositional attitude is a mental state held by an agent with respect to a

representation that bears some truth value. Importantly, propositional attitudes are readily expressed linguistically in English with verbs which are followed by complement clauses headed by the word “that”. These complement clauses are generally further sub-categorized as content clauses. Among the English words characterized by these qualities, think, know, and believe are some of the most common. Such verbs were understood by early scholars of mindreading to instantiate in language the representations used by organisms to navigate through and make decisions about their environs (Dennett, 1978; Fodor, 1992; Perner, 1988). Consequently, when researchers began to consider the circumstances under which an agent could be said with certainty to understand its representations of the world as distinct from those of other agents, the concept of mental states became synonymous with propositional attitudes (Apperly, 2008; Baron-Cohen, 1997b, 1997a; Baron-Cohen et al., 1997; Leslie et al., 2004), bringing along with it attention to these mental state verbs. This focus elided a number of phenomena that appear intuitively to be mental states which were nevertheless left unexamined for many years, including emotions, perceptions, and intentions (Bugnyar et al., 2016; Golan et al., 2007; Harrigan et al., 2018; Hughes & Dunn, 1998; Stewart et al., 2019; Trueswell et al., 2016; Turner & Felisberti, 2017).

Though the elision of these other categories may represent an oversight with consequences for a complete empirical accounting of the relationship between LR3PMS and mindreading, there is sufficient evidentiary grounding to believe both that the distribution of semantic categories is largely the same across languages (B. Fehr & Russell, 1984; Goddard, 2010; Gray et al., 2007; Haspelmath, 2010; Jackson et al., 2019) and that mental state verbs which take content clauses as complements bear a special, causal relation to mindreading (L. Bloom et al., 1989; Gleitman, 1990; Papafragou et al., 2007; Shatz et al., 1983; Ünal & Papafragou, 2018). These facts may mitigate concerns about the exclusion of other cognitive phenomena by suggesting first that a coding scheme which focuses only on mental state verbs

would constitute a kind of systematic, as opposed to random, error. Because there exists a degree of universality in the set of semantic categories represented across languages, it may be reasonable to think that the exclusion of a category from coding would impact the set of languages sampled equivalently. Though it is unknown whether the frequency of a given semantic category varies across languages for a fixed subject of speech (and determining as much is a goal of the present research), these data suggest such concerns may be minimal. Secondly, if it is truly the case that mental state verbs which take content clauses as complements bear a unique causal relationship to mindreading, then exclusion of other kinds of words which refer to other cognitive phenomena may not constitute a limitation on the ability of the present research to address its substantive questions.

The present research pursues this approach by drawing upon empirical data reported by Wellman and Estes (1987) aimed at determining whether children's production of mental state verbs constitute genuine instances of mental state reference. Per the authors, words like "think" and "know" are often used conversationally, as illustrated in sentences such as "You know what?" (to get an interlocutor's attention) and "I think we should get started" (to soften a command). These differ from what they consider genuine instances of mental reference, as in a sentence like "John knew where the item was." In the course of characterizing a prior longitudinal study of early childhood speech (Shatz et al., 1983), the authors found that over 95% of the mental state references children produced were mental state verbs, with a specific emphasis on the following eight: know, think, mean, forget, remember, guess, pretend, and dream. This set of words and their associated conjugations were selected for coding in the present study due to their coherence with early theorizing about the nature of mindreading as a universally human trait characterized by the ability to represent one's own and others' propositional attitudes. Failure to observe universality in the frequency with which such LR3PMS are made might point to flaws in current theories of mindreading. Many have relied on intuitions

and presumptions that may themselves be artifacts of a culturally-situated philosophical tradition the generalizability and universality of which may be overstated.

As stated earlier, the current research cannot answer causal questions in a direct way, nor can it disentangle the effect of language and culture. The only question to which it can provide direct evidence of an answer is whether the rate of LR3PMS varies across three distinct cultural-linguistic samples for a limited, albeit theoretically motivated, set of lexical items. Nevertheless, the current research may point toward answers for some of these other questions. For example, the lack of independent variation between language and culture in the present research may not necessarily pose a problem for narrowing down the range of plausible models. However, the extent to which this is true may be dependent upon the theoretical concern at hand. That is, the current research cannot answer narrow questions about whether variation in language correlates with or causes variation in mindreading. If instead the concern is about whether mindreading varies across cultures, and if language is understood to be just one of many media through which culture is made manifest, then disentangling these two phenomena is less problematic. While the results of this study are also incapable of addressing these kinds of causal questions (namely, whether variation in culture can cause variation in mindreading), merely showing that there exists variation in speech about the mind is suggestive initial evidence of cultural or linguistic phenomena influencing mindreading. Critically, if no variation is observed in LR3PMS across linguistic-cultural samples, then there is suggestive initial evidence to discount models which emphasize direct causal relationships between mindreading and language as well as between culture and language.

Methods

Participants

The participant population employed in this study was collected according to and constituted by the same population described in Chapter 2 – “General Methods” of this

dissertation. Refer to Chapter 2 for a more detailed description of the population characteristics as well as a substantive accounting of the strategy according to which they were recruited.

Materials

The materials used to generate the dataset to be analyzed in the present study can be found in Chapter 2 – “General Methods” of this dissertation. Refer to Chapter 2 for a thorough inventory of the stimuli and software platforms according to which the data were generated.

Procedure

The present study represents just one among many potential procedures according to which the corpora of speech samples collected as part of this dissertation and the Geography of Philosophy Project more broadly may be coded. For a complete description of the protocol used to generate the Mandarin Chinese, Moroccan Arabic, and American English corpora analyzed here, please refer to Chapter 2 - “General Methods” of this dissertation.

Coding

A set of target lemmas (i.e., root word forms) were derived from Wellman and Estes (1987) and translated by coders into each of the target languages. This set of lemmas constituted eight English mental-state verbs: “know”, “think”, “mean”, “forget”, “remember”, “guess”, “pretend”, and “dream.” Coders were tasked with identifying the single most direct translation of each term into their target languages, ignoring synonymous or near-synonymous terms. Initial efforts revealed complications with respect to the cross-linguistic commensurability of these terms. Although the semantic and conceptual scope of the referents indexed by these terms in English may seem like natural kinds, they are in fact just one of many possible mappings between the semantic-conceptual referent space and the lexicon of a language. Distinctions within this space glossed with a single lemma by one language may be mapped into two or more grammatically distinct lemmas in another. As an example, the infinitive English lemma “know” can be used to refer both to instances in which there is knowledge of facts and to

instances in which there is familiarity with individuals. In contrast, Spanish maps these two sorts of knowing into separate infinitive lemmas – “saber” and “conocer,” respectively. These terms are not synonyms, but instead distinct lemmas mapping the same semantic-conceptual space as the English word “know” in ways that are grammatically meaningful to fluent Spanish speakers. Therefore, a complete translation of the term “know” from English into Spanish cannot be accomplished without counting both “saber” and “conocer”.

Recognizing this complication, coders in each language were asked to identify whether candidate translations of each English term could be understood as mere synonyms or distinct lemmas that carved up the semantic-conceptual scope of the English term in incommensurable ways. Where the latter was true, multiple such lemmas were permitted (so some languages had more than eight mental state verbs coded). Once lists of translated lemmas had been generated for each language, coders were provided with the Dictionary spreadsheet corresponding to their first language and asked to identify all unique word types that represented inflected (declined or conjugated) forms of the lemmas in their translated list. Coded Dictionary spreadsheets were then used in conjunction with a custom Python script to automatically code the Raw Data File. Coders were next provided with copies of the coded Raw Data spreadsheet and asked to review all coded word tokens. This review was focused on determining whether the token constituted an LR3PMS (i.e., that of a character depicted within one of the video stimuli). After both coders for a given language had completed their review of the coded tokens in the Raw Data spreadsheet, interobserver reliability analyses were performed on their respective encodings using intra-class correlations. Where ICC coefficients were above 0.9 (indicating excellent agreement), data for a given language were not subject to additional processing and were ready to be analyzed. Where ICC coefficients were below this value, coders were asked to independently review the data points on which they disagreed and determine which, if any, should be included. Newly coded items in each of the Raw Data spreadsheets were highlighted

and were once again reviewed by both coders to ensure they constituted LR3PMS. Upon completion of this second review, interobserver reliability analyses using intra-class correlations were run again. This review process was repeated no more than two times per language before all ICC coefficients were above 0.9.

Data Analysis

As the finalized data to be analyzed constituted discrete counts of LR3PMS per transcript, which were themselves sets of non-independent observations collected within participants, an analytic strategy was adopted, according to which the data were first analyzed using pure random-effects, or variance components Poisson models. All models were specified using the 'glmer' package in R. The dependent variable of interest was the count of LR3PMS produced in a given transcript (*LR3PMS*), while the predictor variables of interest were the identity of the participant who produced the transcript (*Participant ID*), the video to which the description contained in the transcript corresponded (*Video ID*), and the field site from which the participant who produced the transcript was recruited (*Field Site*). Because a core question of the present study was whether *Field Site* determined a substantial degree of variation in the production of LR3PMS, a model comparison procedure was implemented wherein the goal was to observe whether the conditional modal estimate of the random effect of *Field Site* changed significantly across models featuring other predictors. Across all models, a random effect of *Participant ID* was included due to wide variation in mean transcript length (number of words uttered) across participants, which also correlated with the count of LR3PMS. Furthermore, *Participant ID* was nested within *Field Site* to capture the sampling structure of the data. Next, the dataset was split according to *Vid ID*. Separate general linear models were run for each of these nine datasets to predict counts of LR3PMS as a function of *Field Site.* This approach permitted the exclusion of both *Video ID* and *Participant ID* from the model, as there were no repeated measures from individual participants within each dataset and each dataset

corresponded to a unique *Video ID.* These models generated predicted counts of LR3PMS to compare against observed counts and to make predictions about the expected number of counts across a range of transcript lengths for participants from each of the three field sites.

Results

Variance Component Model Comparison

A total of three separate variance component models, or pure random-effects models, were run in order to determine the proportion of variance accounted for by each predictor in the model, as well as to derive the estimated variance in each predictor. Three models were run in order to determine the effect the predictors had on each other with respect to the estimated variance attributable to each factor. All models were run as Poisson regression models using the 'glmer' function of the 'lme4' package (Version 1.1-35.3) in the R statistical programming language (Version 4.4.1). The models, in order, were specified as follows:

1. *LR3PMS ~ (1|Video ID*Field Site) + (1|Video ID) + (1|Field Site) + (1|Field Site / Participant ID) + Offset(Log(Total Words Uttered))*
2. *LR3PMS ~ (1|Video ID*Field Site) + (1|Video ID) + (1|Field Site) + (1|Participant ID) + Offset(Log(Total Words Uttered))*
3. *LR3PMS ~ (1|Video ID) + (1|Field Site) + (1|Participant ID) + Offset(Log(Total Words Uttered))*

The terms in these models indicate that the count of LR3PMS was modeled as a function of random intercepts for each level of *Video ID* (specified as *(1|Video ID)*), random intercepts for each level of *Field Site* (specified as *(1|Field Site)*), random intercepts for each level of the interaction between *Video ID* and *Field Site* specified as *(1|Video ID*Field Site),* and random intercepts for each level of *Participant ID* (specified as *(1|Participant ID),* each of which was also allowed to vary across *Field Site* to account for the nesting structure between these two variables (specified as *(1|Field Site/Participant ID)* though equivalent to *(1|Participant ID)* as

will be illustrated in the following two sections). Additionally, an offset term was included to account for the fact that the greater the number of words uttered, the greater the number of exposures within which a LR3PMS could occur.

Variance Component Model 1 (VCM 1)

VCM 1 was fit to examine the role of *Video ID, Field Site,* and their interaction simultaneously. VCM 1 fit was evaluated using the AIC, or Akaike Information Criterion (AIC = 1972.5) and the BIC, or Bayesian Information Criterion (BIC = 199.3). The log-likelihood of the model was also reported (log-likelihood = -981.2). Variance estimates and standard deviations for each predictor in the model were assessed to determine the variability the random effects captured, the results of which can be found in **Table 2**. Predictors are referred to by their variable names in plain English rather than using the syntax of the model to which they corresponded. The variables for which variation was greatest between levels were *Video ID* (var = 0.56473, sd = 0.715) and the interaction between *Video ID* and *Field Site* (var = 0.4947, sd = 0.7034). Crucially, variance estimates for the remaining variables exhibited two features of note. First, *Participant ID* nested within *Field Site* (var = 0.12428, sd = 0.3525) was substantially greater than that of *Field Site* alone (var = 0.02472, sd = 0.1572). Second, both of these values were substantially lower than either *Video ID* or the interaction between *Video ID* and *Field Site.* Intraclass Correlation Coefficients were calculated for each random effect to determine the proportion of the total variance explained by each. 10.28% of the total variance explained by the model was attributable to *Participant ID* nested within *Field Site*, 46.73% was attributable to the interaction between *Field Site* and *Video ID*, 40.94% was attributable to *Video ID*, and only 2.05% was attributable to *Field Site*. **Figures 2, 3,** and **4** illustrate the conditional modes of the random intercepts estimates with 95% confidence intervals for each variable in the model.

As can be seen in **Figure 2**, the 95% confidence intervals for all three field sites overlap substantially with each other and include zero, indicating that each level of *Field Site* differs

neither from the others nor from the grand intercept estimate. In **Figure 3**, the 95% confidence intervals for each of the nine video stimuli indicated that the only video stimulus which reliably differed from the grand intercept estimate was False Belief, though the 95% confidence interval for Mate Guarding overlapped with zero only slightly. Confidence intervals on the conditional modal intercept estimate for Mate Guarding overlapped with those of every other video, while those for False Belief video reliably differed from those of Dangerous Animal and Cooperation. These results suggest that at least one, though possibly two videos tended to elicit a greater number of LR3PMS than average. These results also support the conclusion that the count of LR3PMS in transcripts describing False Belief was reliably higher than in transcripts describing Dangerous Animal and Cooperation.

In **Figure 4A,** none of the conditional modal estimates of the random intercepts for *Video ID* reliably differed from zero or from each other among participants recruited from China. The same can generally be said of participants recruited from the United States, with the exception of the conditional modal estimate of the intercept for Cooperation which was found to be reliably below average. In contrast, three of the conditional modal estimates of the random intercepts for *Video ID* differed reliably or nearly reliably from zero among participants recruited from Morocco. Here, estimates for False Belief and Prestige were higher than the grand intercept estimate, while nearly reliably lower for Sickness. **Figure 4B** presents the same data as **Figure 4A** grouped by *Video ID* on the y axis and illustrates that the conditional modal estimates of the random intercepts for each country do not differ from each other across any of the video stimuli.

Variance Component Model 2 (VCM 2)

Given the results of VCM1 which suggested only a minimal fraction of the variability in the data was attributable to *Field Site*, VCM tested whether nesting *Participant ID* within *Field Site* misattributed variation in *Field Site* to *Participant ID.* As such, VCM 2 replicated the overall

structure of VCM 1 while treating *Participant ID* as non-nested. The results of VCM 2 were the same as VCM 1, tables and figures for which can be found in Appendix C.

Variance Component Model 3 (VCM 3)

A primary finding of VCM 1 was that nearly 90% of the variance in the data was attributable to the combined effect of *Video ID* and the interaction between *Video ID* and *Field Site*, while almost none of the variance in the data was attributed to *Field Site* alone. While VCM 1 provided evidence to suggest that *Field Site* is a relatively unimportant source of variance in transcript counts of LR3PMS, the potential collinearity of the random effect term for *Field Site* and the random effect term for the interaction between *Video ID* and *Field Site* challenge this interpretation. VCM 3 sought to determine address this limitation. VCM 3 fit was evaluated using the AIC, or Akaike Information Criterion (AIC = 2039.5) and the BIC, or Bayesian Information Criterion (BIC = 2061.0). The log-likelihood of the model was also reported (log-likelihood = -1015.8). Variance estimates and standard deviations for each predictor were obtained to determine the variability the random effects, the results of which can be found in **Table 3**.

With the exclusion of the random effect term for the interaction between *Video ID* and *Field Site*, the ordering of the variables by the amount of variance attributed to them was the same for VCM 3 as it was for VCM 1. The variance estimates were greatest for *Video ID* (var = 0.5296, sd = 0.7277), followed by *Participant ID* (var = 0.1238, sd = 0.3519) and finally by *Field Site* (var = 0.1103, sd = 0.3321). Perhaps unsurprisingly, the amount of variance attributed to *Field Site* did increase relative to VCM 1. However, the increase did not represent a simple transfer from the interaction term to the *Field Site* term, but a complex reapportionment in which a substantial amount of explained variance was lost. While some of the variance attributable to the interaction term overlapped with that of *Field Site,* it appears to be the case that a substantial proportion was uniquely attributable to the particular effects of particular videos within each sample across field sites. Nevertheless, VCM 3 reaffirms the findings of VCM 1

wherein *Video ID* appears to account for a greater degree of variance in the data than does *Field Site.* Intraclass Correlation Coefficients were calculated for each random effect to determine the proportion of the total variance explained by each. 16.21% of the total variance explained by the model was attributable to *Participant ID*, 69.35% was attributable to *Video ID*, and 14.44% was attributable to *Field Site*. While this represents a notable increase from the proportion attributable to *Field Site* in VCM 1, it remains accountable for less variance than *Participant ID* or *Video ID.* **Figures 5** and **6** illustrate the conditional modes of the random intercepts estimates with 95% confidence intervals for each of the variables in the model. The values presented in **Figures 5** and **6** represent simultaneously the extent to which each level of the variable differs from the grand intercept estimate and from each other.

As can be seen in **Figure 5**, the 95% confidence intervals for the conditional modal estimates of the intercept for only one of the three field sites, China, overlaps substantially with zero. Nevertheless, the conditional modal estimates for China and Morocco overlap substantially, as do the conditional modal estimates for China and the United States. Collectively, these results indicate that the conditional modal intercept estimates for the United States and Morocco differ reliably from the grand intercept estimate and from each other, though neither field site differs reliably from that of China which is itself essentially identical to the overall average intercept. In **Figure 6**, the 95% confidence intervals for each of the nine video stimuli indicate that five videos have conditional modal intercept estimates that reliably differ from that of the grand intercept estimate. These videos, in order of the absolute magnitude of difference from the overall intercept, are the False Belief video stimulus (greater than average), the Mate Guarding video stimulus (greater than average), the Cooperation video stimulus (less than average), the Norm Violation video stimulus (less than average), and the Dangerous Animal video stimulus (less than average). Of these conditional modal intercept estimates, Norm Violation, Dangerous Animal, and Cooperation did not reliably differ from each

other, though they did differ from Mate Guarding and False Belief. The conditional modal intercept estimates for these two videos reliably differed from each other. In total, these findings point toward meaningful variation in the extent to which the different video stimuli elicit LR3PMS.

Based on both the fit statistics and the interpretation provided above, VCM 1 represented a better fit to the data, with lower values than VCM 3 across AIC, BIC, and log likelihood scores. Moreover, VCM 1 accounted for a greater proportion of the variance in the count of LR3PMS within transcripts than VCM 3, the results of which were confirmed with a Chi-Square difference test that was highly statistically significant, $X2\ (1, N = 2) = 69.036$, $p < .0001$. This result indicates that the larger model (VCM 1), with a greater number of estimated parameters fits the data more closely than the smaller model (VCM 3). Therefore, subsequent analysis of the data is based upon the results of VCM 1. A consequence of this finding was that the best fit model, VCM 1, attributed very nearly none of the variance in transcript counts of LR3PMS to *Field Site.* This result will be considered in greater detail in the discussion.

Evaluation of Model Fit

Figures 7 and 8 illustrate both the observed and fitted mean counts of LR3PMS for the interaction between *Video ID* and *Field Site* (**Figure 7**), as well as for the main effects of *Video ID* and *Field Site* (**Figure 8**). Notably, the fitted estimates of the mean count of LR3PMS among transcripts corresponding to each of the levels of these predictors were very close to the observed values, further indicating good model fit. Next, a simulated dataset was generated to ascertain the predicted average count of LR3PMS for *Video ID*, *Field Site*, and their interaction when transcript length was held constant. Transcript length values corresponded to the lower quartile (n=40 words), the median (n = 85 words), and the upper quartile (n = 142 words) of observed transcript lengths. The simulated dataset contained 4779 observations corresponding to three transcripts varying in overall length (40 words, 85 words, and 142 words) per participant

(n=177) per video stimulus (n = 9). Field site was left unmanipulated across participants to account for the fact that they could not have been drawn from different sites.

The simulated dataset was fed into VCM 1 and the resulting predictions, with standard errors, were used to produce mean predicted counts of LR3PMS and associated standard errors for each level of the predictors in the model. These results were plotted and can be found in **Figures 9 – 11**. Predicted counts of LR3PMS for the interaction between *Field Site* and *Video ID* across transcripts at the lower quartile value, the median value, and the upper quartile value of transcript length can be found in **Figure 9.** As can be seen most clearly in **Figure 9A**, the confidence intervals for almost every level of the interaction term, at each of the three specified transcript lengths include zero, indicating that predicted counts of LR3PMS are neither reliably different from zero nor are they reliably different from each other. The only levels for which the predicted count of LR3PMS was reliably above zero were for transcripts 85 and 142 words length produced by Moroccan participants while describing the False Belief video stimulus. **Figure 9B** presents the same data organized by *Field Site* on the x axis.

Next, observations were collapsed across *Field Site* to permit observation of the predicted count of LR3PMS for each level of *Video ID* for transcripts 40, 85, and 142 words in length. As can be seen in **Figure 10**, the predicted count of LR3PMS was reliably different from zero for all levels of *Video ID* across all transcript lengths. However, only False Belief and Mate Guarding yielded predicted counts reliably different from each other (with reliably higher predicted counts for False Belief than for Mate Guarding) and reliably of greater magnitude than the seven remaining video stimuli at each of the three pre-determined transcript lengths. Across transcript lengths, Prestige yielded reliably higher predicted counts of LR3PMS than did Norm Violation, Infidelity, Dangerous Animal, and Cooperation, but not Sickness or Dominance. Additionally, predicted counts of LR3PMS for Prestige were reliably lower than those of False Belief and Mate Guarding. In **Figure 11**, observations were collapsed across *Video ID* to

quantify the predicted count of LR3PMS for each level of *Field Site* across transcripts of 40, 85, and 142 words in length. Across all transcript lengths, the average predicted count of LR3PMS in transcripts produced by participants from China, Morocco, and the United States were reliably above zero but not reliably different from each other. Surprisingly, mean predicted counts of LR3PMS were higher for Morocco when holding transcript length constant at the first quartile, median, and third quartile values. This finding represented a departure from the observed (**Figure 8B**) and fitted values (**Figure 8D**) wherein mean LR3PMS counts were lowest for transcripts made by Moroccan participants when compared to Chinese or American participants. To shed light on this contradictory finding, mean transcript lengths were plotted as a function of *Video ID, Field Site,* and the interaction between the two variables. Description of the observed data in this way is presented in **Figures 12 and 13.** A striking difference in the average length of transcripts produced by participants in Morocco relative to participants in China or the United States is especially readily observed in **Figures 12 and 13.** Averaging across videos, transcripts produced by Moroccan participants are reliably shorter than those produced by American participants, and nearly reliably shorter than those produced by Chinese participants (**Figure 13B**). This same pattern holds at least as strongly, if not more so, when broken down by the particular video stimuli to which a transcript corresponds (**Figure 12A**). Additionally, **Figures 12B and 13A** appear to suggest that the rank ordering of video stimuli by mean transcript length is more or less the same across the levels of *Field Site*, indicating that the total number of words uttered may be tracking a property of the video stimuli themselves, such as duration in seconds. Pearson's product-moment correlation was conducted on these data and a small, but highly statistically significant relationship was found between the total number of words uttered in transcripts and the length in seconds of the video to which the transcript corresponded, $r(1587) = .122$, $p < .0001$. Thus, though the mean length of transcripts varied substantially across field sites, these values may have been tracking structural features of the content to which they

corresponded – in particular, the length of the video stimuli. However, subsequent correlation analyses suggested that the length of the video stimulus to which a given transcript corresponded was far more strongly correlated with the count of LR3PMS, $r(1587)=0.302$, $p<.0001$. This correlation was stronger than that observed between the total words uttered in a given transcript and the count of LR3PMS, $r(1587)=0.2562$, $p<.0001$. Consequently, the strongest predictor of the production of LR3PMS may be related to narrative elements of the stimuli as opposed to structural features.

Discussion

Here, I found that when speakers of Moroccan Arabic, American English, and Mandarin Chinese were asked to describe a standardized set of 9 video stimuli depicting naturalistic social interactions, no cross-linguistic differences in the frequency of mental state talk were observed (**Figures 2, 7, and 8**). Further cementing this point, individual differences between participants were found to account for more of the variation in the frequency of mental state talk than were cross-linguistic differences. Notably, however, I also found statistically significant differences in the total number of words speakers of each language uttered such that Moroccan Arabic speakers uttered fewer words on average than did American English or Mandarin Chinese speakers (**Figures 12 and 13B**). Though the languages did not differ in the real frequency of mental state talk, they were nevertheless found to differ in the relative frequency or rate of mental state talk. Consequently, when predicted counts of mental state talk were generated holding transcript length constant, Moroccan Arabic speakers were predicted to produce significantly higher counts of mental state talk when compared to American English speakers (**Figures 9 and 11**). I also found that one of the 9 video stimuli in particular, the False Belief Video, resulted in statistically significantly more mental state talk than all of the others (**Figures 3, 7, 8, and 10**). Critically, though, there were some videos that elicited lower-than-average and higher-than-average amounts of mental state talk among speakers of particular

languages. Among American English speakers, transcripts describing the Cooperation video had lower than average amounts of mental state talk. Among Moroccan Arabic Speakers, transcripts describing the Sickness video had lower than average amounts of mental state talk while transcripts describing the Prestige and False Belief video had higher than average amounts of mental state talk.

These results constitute four notable preliminary findings that may speak meaningfully, if tentatively, to the questions this dissertation aims to address. The first of these three preliminary findings is that speakers of different languages or people across cultural environments do not vary in the frequency with which they talk about mental states when talking about the same topic or subject, as evidenced by the estimated variance attributed to *Field Site* in VCM 1. Though such an interpretation may find support in the literature among those who advocate a universalist view of grammar and language (Fitch et al., 2005; Hauser et al., 2002), it remains unclear to what extent such universalist accounts entail universality in semantic categories or in the conditions that elicit their manifestation in speech (Haspelmath, 2010; Rauthmann et al., 2014). While there appear to be at least some such semantic categories, and some contexts which seem to more or less universally require certain semantic categories, these generally tend to be limited to cases wherein the referents of such semantic categories are invariant across human populations or in cases where there may have been stabilizing selection on the human cognitive or behavioral phenotype (Christiansen & Kirby, 2003; Dunbar, 2004; Scott-Phillips, 2014; Seyfarth & Cheney, 2014). Though mental states may represent a universal semantic category (Avis & Harris, 1991; Goddard & Wierzbicka, 1994; Norenzayan & Heine, 2005; Wierzbicka, 1996), and though the video stimuli employed in the study were intended, as much as possible, to depict situations which themselves may manifest across all cultural and linguistic populations, the significance of the interaction term in VCM 1 may be understood as reflecting the fact that while culture or language does not exhibit a direct effect on the frequency

with which speakers produce LR3PMS, it may provide a lens through which the meaning of some stimulus is understood and through which the appropriate semantic categories are drawn upon when speaking about it. That is, though the semantic category of “think” may be universal, and though the category of “snake” may be as well, what one thinks upon encountering a snake and whether that thought ought to be communicated may depend on the culturally- or linguistically-structured meaning of such a context. Indeed, there is an abundance of data to suggest that both the meaning of and boundaries on categories of mental states exhibit substantial cross-cultural and cross-linguistic variation (Goddard, 2010; Jackson et al., 2019; Matsumoto, 1989). While the neural, physiological, or even psychological phenomena may be universal, how they are carved up may vary substantially across contexts.

The second of these preliminary findings is that the subject or topic of speech is a far stronger determinant of how much mental state talk occurs than language itself, and that some subjects or topics may engender more mental state talk regardless of cultural or linguistic contexts, as evidenced by the estimated variance attributed to *Video ID.* Across the board, the amount of mental state talk was highest in transcripts describing the False Belief Video, followed by transcripts describing the Mate Guarding video, regardless of which language participants spoke (**Figures 7, 8, and 9**). Affirming conclusions drawn from the analysis of the random intercept estimates, predicted counts of LR3PMS in transcripts describing the False Belief and Mate Guarding video stimuli were found to be higher than those of the others, independent of transcript length (**Figure 10**). This same pattern held for the predicted count of LR3PMS in transcripts produced by participants across field sites, wherein no statistically significant difference was found regardless of transcript length. Importantly, however, it may not be the case that categories of scenarios or situations / categories of speech topics and subjects vary in the extent to which they require mental state talk. It may instead be the case that some of the video stimuli simply happened to have more mentalistic content to describe. Given that

these two video stimuli, the False Belief video and the Mate Guarding video, corresponded respectively to an analog of the False Belief Test, which is itself designed to elicit the attribution of mental states (Baron-Cohen et al., 1997) and to a narrative in which a character experiences sexual jealousy, which has been argued to occur not just across human populations but across species (Buunk et al., 1996; Daly et al., 1982), it is perhaps not a surprise that these would elicit higher counts of LR3PMS.

The third of these preliminary findings is that even if some subjects of speech engender more mental state talk than others, regardless of linguistic or cultural context, there may nevertheless be some which can take on or lose a mentalistic framing in a linguistically- or culturally-determined way, as evidenced by the estimated variance attributed to the interaction between *Video ID* and *Field Site*. In effect, then, the extent to which a given subject or topic of speech is discussed through mentalistic terms may itself be a product of cultural phenomena interacting with more universal cognitive substrates. This notion is analogous to the idea of cultural attractors (Sperber, 1996), wherein the likelihood that certain ideas or cultural phenomena emerge across populations is potentiated by their “fit” to the mind. That is, though any idea or cultural phenomenon is possible in theory, some may be more likely to occur given their concordance with the existing psychological or cognitive architecture. If that architecture is itself universal, then so too might the corresponding ideas or phenomena be. Given the finding that both *Video ID* and the interaction between *Video ID* and *Field Site* explained the overwhelming majority of variance in the data, and that two of the nine videos reliably elicited the greatest mean counts of LR3PMS, these particular video stimuli might represent a kind of basin of cultural attraction. In effect, though the other video stimuli varied more freely in their meaning, these two might tend to “fit” the mind in similar ways across diverse cultural contexts..

The fourth, and arguably most important finding to take away from these results pertains to the way in which “mental state talk” was operationalized in the current study. That is, it may not be

suitable to treat “mental state talk” as equivalent to the production of a set of just eight mental state verbs. Though “think”, “know”, “believe”, “remember”, “forget”, “mean”, “dream”, and “pretend” all constitute genuine mental states, these results highlight the fact that the extent to which mental state talk varies across subjects of speech may depend almost entirely on the set of phenomena you decide constitute “mental states”. It is likely for these reasons that the average frequency of mental state talk was effectively zero for 7 of the 9 video stimuli, despite the fact that they had all been designed to be interpretable only if the participant was inferring the mental states of the agents in the videos. One likely explanation is that mental states like thinking, knowing, and believing were simply inappropriate mental states to meaningfully describe the video stimuli other than the False Belief and Mate Guarding videos. A cursory review of the word types produced by American English speakers includes at least 25 tokens each of the “confused”, “want”, “mad”, “notice”, “upset”, “tired”, “jealous”, and “attention” across the corpus. Each of these word types almost certainly constitute mental states, yet they are not captured by this narrow view of mental state talk. Though the former set of mental state verbs identified by Wellman and Estes (1987) has been purported to play an especially important role in the development of mindreading by virtue of their frequency in English, it is hard to imagine that these other terms are doing no such work. Even if they are less frequent, they may still be doing important work in underscoring the occurrence of a broader range of mental states across a broader range of speech subjects.

Despite the consistency of the observed and fitted counts of LR3PMS, a comparison between these counts and the predicted mean counts of LR3PMS yields two divergent conclusions. While all three of these different views (predicted counts of LR3PMS, observed counts of LR3PMS, and fitted counts of LR3PMS) of the data support the conclusion that there exist no statistically significant differences in the frequency of mental state talk across each of the three field sites, the rank ordering of the predicted frequency of mental state talk across field

sites for transcripts of a fixed length (**Figure 11**) is very nearly the inverse of the fitted and observed frequencies of mental state talk, which themselves mirror each other and therefore speak further to the fit of VCM 1 (**Figures 8B and 8D**). This result is one that requires explanation. One possibility is that the divergence between the fitted frequencies of mental state talk and the predicted frequencies of mental state talk for transcripts of a fixed length is an artifact related to the stark difference in the average length of transcripts produced by participants from Morocco relative to participants from China and the United States (**Figures 12 and 13**). While the fact that the mean total number of words in transcripts produced by Moroccan participants is much lower than that of transcripts produced by Chinese or American participants warrants further investigation, these data suggest that there may be cross-linguistic variation in the number of words one needs to utter in order to communicate a given semantic unit. Such an account constitutes a parsimonious, if untested explanation for both the relative invariance in mean counts of LR3PMS across field sites as well as the variation in mean transcript length. Stated alternatively, participants from Morocco may simply require fewer words to convey the same semantic content, a fact consistent with the synthetic morphology of Arabic which permits the formation of words with more complex meaning than might be permitted by the analytic morphologies of English and Mandarin Chinese (Ezeizabarrena & Garcia Fernandez, 2018).

As such, it may not be meaningful to compare or make predictions about counterfactual, fixed word volumes, as was done here when comparing the predicted frequencies of mental state talk. Future research in which transcripts are coded for morphological or semantic units may be able to speak more directly to the question of whether Moroccan Arabic is more lexically “efficient”, using fewer words to express the same semantic content. Alternatively, it may be possible that speakers of Moroccan Arabic were indeed producing less content than American English and Mandarin Chinese speakers. To address this question, future research may focus

on the semantic richness of the descriptions in order to determine whether speakers of different languages vary in the completeness of their descriptions of the video stimuli.

These findings are thus worth treating cautiously, given the narrow and culturally-situated lens through which the analytic targets were selected. The current study sought only to model the frequency of mental state talk using a set of 8 lemmas that have been thought by western psychologists and philosophers of mind to be related to the mindreading capacity in functionally meaningful ways (Dennett, 1978; Fodor, 1992; Gopnik & Astington, 1988). While such a scheme constitutes a strong starting point for inquiry, this list of eight words does not come close to representing the whole of the mental state lexicon in English, to say nothing of the other languages sampled. Less straightforwardly, it is not obviously the case that these lemmas constitute universal categories or categories that do the same conceptual lifting everywhere. As such, these findings may be limited both by the principles according to which coding targets were identified and by the simple fact of insufficient flexibility to capture mental state talk not contained within the list of 8 lemmas.

Conclusion

In this chapter, transcripts generated according to the procedure detailed in “Chapter 2 – General Methods” were coded for lexical references to third-party mental states (LR3PMS) in order to determine whether transcripts generated by participants sampled from three culturally and linguistically distinct populations (first-language speakers of Mandarin Chinese in China, first-language speakers of Arabic in Morocco, and first-language speakers of English in the United States) and tasked with describing nine novel, silent video stimuli varied, on average, in the count of such references. Transcripts were coded using an inventory of eight lemmas that corresponded to some of the most common mental state verbs used in English and which were derived from the early literature on the human mindreading capacity (Wellman & Estes, 1987). Through the comparison of three distinct variance components models, the best fit model was found to attribute the vast majority of variance in the dependent variable to the

Video stimulus to which a given transcript corresponded and to the interaction between the video stimulus to which a given transcript corresponded and the field site from which the participant who produced the transcript was recruited. Almost none of the variance was attributed to the field site alone. Though these data constitute a compelling initial set of findings which may support both universalist and relativist understandings of talk about the minds of others, the findings presented here might actually depend on the fact that such a small set of words was used. Importantly, this narrow set of words was selected deliberately and aimed to see if focusing on just these words which have been given a privileged status among Western scholars leads to erroneous conclusions, such as the estimated number of mental state terms uttered across a range of contexts being close to zero. This result suggests that people may not actually be talking about the mind, when the opposite is likely to be true. Consequently, a more meaningful approach should aim to capture all mental state talk as understood by native speakers. Doing so may thus illustrate that mental state talk happens across a broader range of contexts, and in more or less similar ways despite the linguistic or cultural environment. In the next chapter, I implement a novel and culturally-variable coding scheme on the same corpus of transcripts to determine whether the patterns seen with the Wellman and Estes terms re-emerge when native speakers of each target language are tasked with identifying all instances of LR3PMS as determined by their respective cultural-linguistic frames of meaning.

Chapter 4: Examining Cross-Linguistic Variation and Uniformity in the Production of All Mental State Terms

Introduction

In Chapter 3, a cross-linguistic corpus of standardized and systematically collected speech samples was coded for the occurrence of LR3PMS using an inventory of eight mental state verbs derived from the early literature on mindreading (Shatz et al., 1983; Wellman & Estes, 1987). This inventory consisted of the words, “think”, “know”, “believe”, “pretend”, “mean”, “dream”, “remember”, and “forget”. Statistical models were run wherein the count of LR3PMS in a given speech sample was analyzed as a function of the identity of the participant who produced it (*Participant ID)*, the video stimulus to which the speech sample corresponded (*Video ID*), the field site from which the participant who produced the speech sample was recruited (*Field Site*), and the interaction between *Video ID* and *Field Site.* An offset term was included to account for variation in the length of each speech sample (*Total Length*).

The results from the study in Chapter 3 suggest a lack of difference in the frequency of LR3PMS across field sites as well as the presence of variability in the frequency of LR3PMS across topics of speech, it is nevertheless possible that the approach of coding only eight key words over- or underestimated the frequency of mental state talk. That is, there are substantive critiques of the research tradition from which the coding scheme of Chapter 3 was drawn that highlight meaningful oversights and limit the confidence to be placed in these findings. Beyond issues stemming from the limited definition of mental state talk employed in Chapter 3, the mental states to which they refer represent just a single dimension of the cognitive processes and capacities engendered by mindreading. It thus remains possible that speakers of different languages vary in the frequency with which they produce LR3PMS and that this variation genuinely reflects the extent to which they cognize what is an otherwise universal set of mental state phenomena (Avis & Harris, 1991; Floyd et al., 2018; Goddard, 2010; Huang & Jaszczolt,

2018; Imai & Gentner, 1997; Jackson et al., 2019; Viberg, 1984). In the following sections, I first address some of the other ways in which mental state talk may vary as well as other mindreading processes this earlier coding scheme may have overlooked.

Variation in Mental State Talk Not Captured by Wellman and Estes Scheme

The narrowness of the eight-word Wellman and Estes coding scheme used in my first study might fail to capture the great many ways in which mental state talk might vary across languages—as well as fail to capture many or even most instances when people are referring to others' minds in everyday talk. Languages may vary in the set of concepts used to encode mental states (Barrett et al., 2016; Goddard, 2010; Haspelmath, 2010; Jackson et al., 2019; Proost, 2007), they may differ in the emphasis placed on certain categories or the granularity with which a given category is carved into separate lexemes (Fausey & Boroditsky, 2010; Majid et al., 2007; Meins et al., 2014; Phillips & Boroditsky, 2003; Sutrop, 2001; Thompson & Juan, 2006), and they may vary in the extent to which particular speech genres or subjects require the spoken attribution of mental states to others (Bendix, 1992; Hawkins & Goodman, 2016; Hoenigman, 2015; A. Lindström & Sorjonen, 2012; J. Lindström & Karlsson, 2016). This variation may be related to pragmatic norms or to explicit prohibitions against making them, though these represent just two of many possible accounts of variation in how certain kinds of speech contexts may vary cross-linguistically with respect to mental state talk (Carston, 2004; H. Clark, 1996; Grice, 1975; Robbins & Rumsey, 2008; Sperber & Wilson, 2002). Moreover, speakers of different languages may use other syntactic categories, like adverbs or nouns, to talk about representational mental states, the result of which would be that the coding scheme I employed in Chapter 3 underestimates the average count of LR3PMS among speakers of languages other than English. Alternatively, languages other than English might predominantly talk about mental states through idiomatic or metaphorical expressions, involving fewer words that make explicit reference to mental states but nevertheless constitute implicit mental state

reference (Carston, 2004; Johnson, 1999; Lakoff, 2008; Winner, 1997, 1997). This, too, would underestimate the frequency of such mental state talk among speakers of that language. Such errors might lend themselves to overestimation of the relative frequency of mental state reference in some languages when compared to others, suggesting that the tendency to do varies cross-linguistically. Even if the exclusion of other kinds of mental states did not constitute threats to the validity of Chapter 3's findings, they might still underestimate the use of mental state speech if adult speakers in the sample used a broader range of epistemic mental state terms synonymous with but not included among the eight terms in the coding scheme.

While translations of the eight terms used in Chapter 3 were generated for each of the languages sampled, there are also reasons to be skeptical of the cross-linguistic equivalence of these terms with respect to their usage. As an example, a language may have direct translations of both "remember" and "recall," but the term for "recall" may be used in the same way that "remember" tends to be used in English. If such a case were to hold, the coding scheme employed in Chapter 3 would fail to capture genuine instances of comparable LR3PMS. Alternatively, if the semantic field of a given English-language lemma like "know" is covered by two terms in another language, as is the case in Spanish and for which the translation of "know" includes both "saber" (to know something) and "conocer" (to know someone), failure to include one or the other of the two terms could lead to a variety of errors in the estimated frequency of LR3PMS. If only a single term, like "saber", is included in the coding scheme, then no matter the result there is reason to be skeptical. Failure to include both terms means that a lower frequency of LR3PMS among Spanish speakers, relative to English speakers, could be attenuated when both terms are accounted for. Alternatively, no difference might be observed when accounting for both. Though steps were taken to limit the possibility of such cases, it is only by coding for all mental state terms that a greater degree of certainty can be approached that any difference, or lack thereof, constitutes a credible effect.

There are, in short, a variety of phenomena through which cross-linguistic variation in mental state talk might manifest. Though the extant literature on cross-linguistic differences in talk about the mental states of third parties is quite small, most studies to date have sought to quantify the frequency of LR3PMS (Devine & Hughes, 2019; Hansen et al., n.d.; Ruffman et al., 2002). This approach takes as a given broad cross-linguistic uniformity in the concepts and representational structures which support mental state attribution. Under this view, cross-linguistic variation in LR3PMS is thought to be driven by variation between individuals, subjects of speech, and cultural-linguistic context. The variation present at each of these levels is thus neither variation in the underlying suite of mental state concepts nor variation in the efficacy with which mental state attribution functions. Rather, it can be thought of as variation in the deployment and linguistic manifestation of a universally human mindreading architecture.

In the decades since the early research on mindreading, which cast it primarily through the lens of representation, meta-representation, and propositional attitudes, a growing number of scholars have argued that mindreading is in fact a multi-dimensional construct that processes many kinds of information including but not limited to epistemic representational states (Apperly & Butterfill, 2009; Castelli et al., 2000; Stewart et al., 2019; Tomasello & Carpenter, 2007). Among the mental states to which increased attention has been paid are emotions (Gendron et al., 2014; Golan et al., 2007; Stewart et al., 2019) and perceptions (Carpenter et al., 1998; Castelli et al., 2000; Gray et al., 2007; Hare et al., 2000). This is not to say that earlier research on mindreading ignored these subjects, but the extent to which they have been theorized as core components of the mindreading system was more minimal (Flavell et al., 1981; Matsumoto, 1989; Tan & Harris, 1991; Tomasello, 1988; Tomasello & Farrar, 1986). As such, another limitation of the coding system used in Chapter 3 is that the verbs in its coding scheme are all mental states that have been characterized by some researchers as “cognitive” or “epistemic” (Bretherton & Beeghly, 1982; Devine & Hughes, 2019; Leslie, 1987; Papafragou et al., 2007;

Ruffman et al., 2002). Crucially, it is not known whether languages differ in how richly they make use of or refer to non-epistemic mental states (Gray et al., 2007; Kulke et al., 2019; Turner & Felisberti, 2017; Weisman et al., 2017). As such, it is unclear to what extent any observed cross-linguistic differences or similarities in talk about cognitive/epistemic mental states alone, as reported in Chapter 3, entail corresponding differences or similarities in talk about other mental states. The absence of these items from the coding scheme in Chapter 3 limits the conclusions that can be drawn from the findings presented therein. Lastly, languages may differ with respect to the ways in which they lexicalize distinct mindreading phenomena, with some manifesting not as distinct lexemes but as morphological elements that modify non-mentalistic root words, as can be seen with the existence of obligatory morphological marking of evidentials in some languages (Papafragou & Li, 2001; Tosun et al., 2013). A coding scheme that does not account for such phenomena may underestimate the frequency of such LR3PMS.

The Present Research

Even if the quantification of the frequency of LR3PMS constitutes a legitimate approach to understanding whether there exist systematic cross-linguistic differences in the ways adults talk about the mental states of third parties, the coding scheme employed in Chapter 3 still exhibits limitations that necessitate further investigation. Here, these earlier findings are built upon and strengthened by analyzing the same dataset with a coding scheme that captures a wider breadth of lexical items used to refer to a wider range of third-party mental states. Given the complexity of the relationships between the outputs of the various mechanisms of the mindreading system, the inputs and outputs of the linguistic system, and the role of communicative norms in shaping the realization of LR3PMS, it is unknown if the patterns observed in the preceding chapter would hold for talk about other kinds of mental states, like third-party emotions or percepts. At present, there is no evidence to suggest that the linguistic outputs of these kinds of mindreading are yoked to each other. As such, the fact that Chapter 3

found speakers of Mandarin Chinese produce LR3PMS with the highest frequency does not guarantee that this pattern will hold when accounting for all mental state terms. Even if speakers of Mandarin Chinese continue to show the highest frequency of cognitive or epistemic LR3PMS when a more inclusive coding scheme is used, they should not necessarily be expected to produce the highest frequency of third-party emotion or perception terms as a consequence. Thus, there are good reasons to broaden the coding scheme to capture all mental state terms, and not just the eight verbs coded previously, and to see how analyzing a much larger set of mental state words might change the conclusions that can be drawn about mental state talk across languages.

Here, I code for a set of mental state terms which have the property of being transient internal states (states as opposed to, e.g., traits), the experience of which manifests in an individual's conscious awareness. This definition is much broader than some that have historically been used in the literature on mindreading and language, as well as the definition employed in Chapter 3 (Booth et al., 1997; Callaghan et al., 2005; Kristen et al., 2014; Pinto et al., 2017; Ruffman et al., 2002; Tardif & Wellman, 2000; Ünal & Papafragou, 2018). Defining mental states as such allows a broader net to be cast when considering possible ways in which mental state talk may vary cross-linguistically and thus provides a mechanism through which to address some of the limitations of Chapter 3. A strategy leveraging the knowledge of first-language speakers of the target languages was implemented in identifying lexical items that could be glossed as mental state terms. First-language speakers of the target languages were recruited as coders, and an iterative process of coding was used to arrive at an inventory of lexical items that could constitute mental state references. Using this inventory as a template, transcripts of participant video descriptions were automatically coded for all instances of the items contained in the inventory. Coders then identified in the transcripts those instances that constituted third-party mental state references and removed erroneously included false

positives. These data were then used to run a series of random effects models wherein *Participant ID, Video ID, Field Site,* and the interaction between *Field Site* and *Video ID* were included as varying intercepts terms.

The present study also allows me to see whether the variation in LR3PMS across video stimuli changes when I expand beyond belief-like mental states. It is possible that by broadening the scope of mental states, the amount of variation attributable to any one level of *Video ID* may be attenuated. In summary, this chapter permits examination of the extent to which variation in the production of LR3PMS is attributable to behavioral variation across individuals, variation across subjects of speech in the extent to which they elicit mental state references, and variation across field sites in the mean frequency with which speakers recruited from particular sites tend to produce LR3PMS. Through the use of a diverse range of stimuli, I will be able to determine whether the identity of an individual, the subject of their speech, or the cultural-linguistic milieu from which the individual is drawn exert substantive influence on their production of mental state talk. Through this novel coding procedure, I will be better equipped to address the outstanding conflicts that exist between the developmental psychological and linguistic anthropological research on mindreading, the production of mental state language, and the causal relationships between these two phenomena across development.

Methods

Although the findings documented in the preceding chapter provide preliminary evidence of cross-linguistic similarity and variation in the rate of LR3PMS across cultural and linguistic contexts, firm conclusions are hard to draw given the limitations of the coding scheme employed. Given the narrow band of mental states encoded by the set of lemmas derived from Wellman and Estes (1987), the methods employed in Chapter 3 could underestimate both the amount and kind of variation in mental state talk across the field sites sampled. By examining a broader range of mental state terms and concepts, otherwise unobserved variation may be

revealed, shedding more definitive light on the nature of cross-linguistic and cross-cultural variation in mental state talk. In short, the focus on cognitive and epistemic terms characteristic of earlier research on mental state talk may have arrived at erroneous conclusions about the frequency of LR3PMS across cultural and linguistic contexts. It is this state of affairs the present research aims to address.

Participants

The participant population employed in this study was collected according to and constituted by the same population described in Chapter 2 – “General Methods” of this dissertation. Refer to Chapter 2 for a more detailed description of the population characteristics as well as a substantive accounting of the strategy according to which they were recruited.

Materials

The materials used to generate the dataset to be analyzed in the present study can be found in Chapter 2 – “General Methods” of this dissertation. Refer to Chapter 2 for a thorough inventory of the stimuli and software platforms according to which the data were generated.

Procedure

The present study represents just one among many potential procedures according to which the corpora of speech samples collected as part of this dissertation and the Geography of Philosophy Project more broadly may be coded. For a complete description of the protocol used to generate the Mandarin Chinese, Moroccan Arabic, and American English corpora analyzed here, please refer to Chapter 2 - “General Methods” of this dissertation.

Coding

For this analysis, a corpus of narrative retellings of video stimuli was coded with a novel methodology to capture all putative tokens of LR3PMS. This methodology leveraged the emic knowledge of first-language speakers of English, Moroccan Arabic, and Mandarin Chinese to identify all types of lexical items that constitute plausible mental state terms and to then assess

tokens of those lexical types in their original speech context. Bilingual students whose L2 language was English and whose L1 language corresponded to one of the three target languages were recruited to code the data. These coders were provided with a definition of a "mental state" to guide their work, the specific details of which can be found in the following section. Using this definition, coders processed the data according to the steps specified in Chapter 2 – "General Methods". Components of the coding procedure particular to the current study are described in subsequent sections of the current chapter.

Definition of a mental state. Coders were provided with the following definition of a "mental state", in English, to guide their work:

1. Mental state terms are "words that describe something someone is experiencing internally over the course of the video."
2. Instances of those words that refer to the participants themselves or to the experimenter are not to be coded, as they do not constitute third-party mental state references, but first- and second-party, respectively.
3. Only instances of those words referring to third parties (i.e., characters in the video or absent others) are to be coded.
4. While mental states can often be described using phrases, metaphors, idioms, and expressions, I am interested in only those mental states that can be expressed using a single word given the focus of the present study on LR3PMS.
    a. For example, items like "flabbergasted" and "unknown" constitute strong candidates for mental state terms.
    b. In contrast, items like "having a good time", and "losing her marbles" do not.
5. A mental state can be further defined as a transient condition or episode of the body that takes place internally, the occurrence of which manifests in an individual's conscious awareness.

    a. Those transient conditions or episodes that occur without conscious awareness, like the contraction of smooth muscle in the gut, do not count.

6. Actions that merely index or point to those internal episodes that manifest in an individual's awareness also do not count.
    a. For example, "smile" should not be coded as a mental state because that is an action, even though it very likely indexes a related internal episode of conscious awareness.
7. The items to be coded must explicitly refer to those internal episodes of conscious awareness.
8. The items to be coded should apply only to agents (and not situations, objects, etc.) and should refer explicitly (and not implicitly) to the internal experience of the agent. Even if a word that could be glossed as referring to mental states is used to describe an object or situation, it should not be coded. To illustrate,
    - Agents vs. Objects
        i. **Code this:** *This situation is making her **angry**.*
        ii. **Do not code this:** *This situation is **infuriating**.*
    - Implicit vs. Explicit Reference to Internal States
        i. **Code this:** *It **looked to him** like they were having a serious discussion.*
        ii. **Do not code this:** *It **looked** like they were having a discussion.*

Resolving Coding Disagreements. After all coders for a given language sample had completed reviewing their Raw Data files to the lead experimenter, inter-rater reliability analyses were conducted. Using established guidelines for the magnitude of Cohen's kappa, the necessity of additional data processing was determined as a function of the magnitude of the resulting kappa statistic. Specifically, inter-rater reliability indicative of sufficient agreement between coders had been obtained if the kappa statistic was observed to be above 0.60, as

suggested by Cohen (1960). This value differed from that of the criterion in Chapter 3 owing to the fact that variation could exist in both the lexical item types identified in the Dictionary File and the specific tokens coded in the Raw Data files. Consequently, none of the coders reached this level of agreement outright, and all languages sampled underwent at least a second round of review to identify the source of discrepancies and resolve disagreements between each of the coders for a given language.

Extracting lemmas from Raw Data files and identifying conjunction and disjunction of lemma sets. After performing inter-rater reliability analyses and determining the need for further review of Raw Data file codes, the set of lexical item types coded as third-party mental state references in each of the Raw Data files was extracted and the disjunction of the two sets was produced. The disjunction of the set represented those items that were identified by one, but not the other, of the two coders in their Dictionary Files. To resolve discrepancies across the disjunction of the lexical item types, it was presented to each coder as a list. Coders for each language were told that the list contained those lexical item types coded in their Dictionary Files about which there had been disagreement. Knowing that these items had been coded by at least one of the two coders, they were tasked with reviewing and, where appropriate, recoding those items as constituting plausible mental state terms. These reviews served to restructure the set of lexical item types in their Dictionary Files. This process thus allowed coders to remove those lexical item types from their own Dictionary Files they had erroneously elected to include, as well as to include those lexical item types from the other coder's Dictionary File that they had erroneously failed to include in their own. Each coder performed this review and upon completion, their newly updated Dictionary Files were used to modify the Raw Data Files accordingly. Newly added tokens were highlighted in yellow to permit easy identification of those that now required review in context, as well as to prevent unnecessary review of items they had evaluated in earlier iterations of the coding procedure.

Coders reviewed the highlighted items according to the steps detailed earlier, after which they were subject to a second round of inter-rater reliability analyses. Agreement between coders was found to be good to excellent for all three languages (Cohen, 1960; McHugh, 2012).

Generating Translations of Coded Items for Cross-Linguistic Comparison

Next, all unique lexical item types remaining in the Raw Data Files for each language were extracted and merged into a single list of words to be translated into English. Each coder was given this list and tasked with providing English-language translations of the lexical item type. Working independently, coders provided their own translations of the lexical item types in the list. For each language except English, coders' lists of translated lexical item types were compared in order to standardize the translations for each entry. Translations that matched exactly were left unaltered and thus considered "standardized." Entries for which the translations did not match exactly were first examined for the presence of spelling errors, formatting differences, or additional words that might have obfuscated otherwise identical translations. Where such variations were identified, translations were corrected to match and were subsequently treated as "standardized". Remaining entries for which the translations were mismatched were compared to determine the overlap of their semantic scope. For example, one coder working in Spanish may have provided a translation of the word "asustado" as "scared" while another may have provided "afraid." In these cases, a coin was flipped to determine which of the two translations would be treated as the "standardized" form. This removed bias and permitted the same diversity of lemmas in the data sets for which translations were necessary as was present in the English data set.

Translations for which the semantic scope of the entries did not overlap were reviewed by the lead experimenter to select the more appropriate of the two translations as the "standardized" form or to triangulate a new "standardized" translation, the semantic scope of which split the difference between translations. To do so, the lead experimenter first read each

of the translations. Then, the target word was entered into Google translate to determine whether the most strongly suggested translations aligned more closely with one or another of the translations. In cases where Google translate suggested only a single English translation that matched one of the two coder-provided translations, this term was treated as "standardized". In cases where Google translate suggested both as top candidates, a coin was flipped to determine which of the two translations would be treated as "standardized". In cases where Google translate suggested both terms but there existed a disparity between the two in terms of the strength or frequency of the translation, a weighted coin was flipped to determine which of the two terms would be treated as standard. Finally, in cases where Google translate suggested neither term or suggested both but weakly, the semantic scope of the top candidate translation was compared to those of the coder-provided translations. If, according to the subjective qualitative judgment of the lead experimenter, this term was found to be a reasonable compromise between the two translations and consistent with the goal of the coding procedure, it was treated as "standardized". If, however, it was not found to represent a reasonable compromise or if it was found not to represent a term that could plausibly be used to refer to mental states, a novel term was generated that fit these criteria according to the qualitative, subjective judgment of the leader experimenter. This term was treated as "standardized".

Lemmatizing Translated Items for Cross-Linguistic Comparison

Once all of the pairs of translations had been standardized, they underwent lemmatization, or conversion into a more basal, root, or stem form of the word. The goal of lemmatization in this case was to provide an overarching label for all inflections of a given term, the core semantic content of which was the same. While this process can be straightforward for word roots or stems that exhibit a narrow range of possible inflections, the picture as a whole is substantially more complex. It is often the case that the semantic scope of a particular inflection of a word root differs from that of the root itself, as well as from those of the majority of its other

inflections. In such cases, it is unclear whether to treat this inflected form as a separate lemma, and thus a separate word root, or to treat it as just another one of the possible inflections of the more basal form from which it is derived.

To illustrate, consider the word “caring,” to be understood here as a participle. A participle is a nonfinite verb form that has some of the characteristics and functions of both verbs and adjectives. In essence, it is a word derived from a verb and used as an adjective (e.g., “a caring teacher”). In this particular case, the participle “caring” is derived from the verb “care”. When used in its most frequent colloquial contexts, the semantic scope of the more purely verbal inflections of “care” captures the extent to which some object matters to the subject of an utterance, as in the cases “I don’t care that my shoes are scuffed” and “She cares a lot about renewable energy”. In contrast, although “caring” represents a possible, strictly verbal inflection of care (albeit one whose semantic scope differs from that described above, as in the sentence “They spend a lot of time caring for their elderly parents”), it is more commonly used in its participle form, the semantic scope of which captures an individual’s interpersonal warmth, generosity, and concern for others (i.e., “He is a very caring person”). Abstract of the additional sentential context, it is not possible to determine from the form of the inflection alone whether the occurrence of “caring” among the translations provided by coders should be understood as a gerund, a participle, or a progressive tense inflection of the infinitive “to care”. A further complication that builds on this core theoretical challenge can be found in cases where gerunds, participles, or other such inflected forms of the infinitive take further inflections. For example, the participle “caring” can take the adverbial suffix and undergo further inflection into “caringly”, as seen in the sentence “He looked at them caringly”. As above, it is unclear from first principles whether “caringly” should be lemmatized to “caring”, in which case “caring” represents a lemma distinct from that of “care”, or if “caringly” should be lemmatized to “care”, in which

case all inflections of the infinitive are subsumed under the same umbrella, regardless of the fact that their common use in language and their semantic scopes are not strictly comparable.

While the preceding discussion pertains primarily to differentiating between inflections of verbs that operate themselves as verbs and inflections of verbs that serve other grammatical functions, as well as to the challenges such differentiation poses to lemmatization, these difficulties are not the unique purview of verbs alone. There exist cases in which inflected forms of nouns and adjectives may also present difficulties in lemmatization with respect to capturing the appropriate semantic scope of the target translation. Given the challenges an approach sensitive to all of the above concerns presents, lemmas could be assigned to each translation individually, thereby ensuring their accuracy but increasing labor demands, or they could be assigned systematically, thereby risking the introduction of erroneous semantic glosses but reducing labor demands. A strategy that balanced these tradeoffs was implemented.

First, a list of the most common English inflectional and derivational morphemes was acquired. Then, each standardized translation was lemmatized according to the following rule. If the standardized translation contained one or more derivational or inflectional morphemes represented in the list, and removal of the maximally derived morpheme did not change the semantic scope of the word with respect to its mentalistic content, then the maximally derived morpheme was dropped. This process was repeated for each standardized translation until either all derivational and inflectional morphemes had been removed from the word root or until all morphemes that could be removed without changing the semantic scope of the root with respect to its mentalistic content had been removed. This minimally morphologically derived form of the translation thus served as the lemma for that word. For example, if the standardized translation of a term from Mandarin Chinese was “embarrassingly”, the suffix “-ly” would be dropped, producing the word “embarrassing”. As this word contained the inflectional morpheme “-ing”, the dropping of which would produce the word “embarrass” and the semantic scope of

which still captured the mentalistic content of the original translation, it too could be dropped, yielding the word “embarrass”. As this could not undergo any further morphological reduction, “embarrass” served as its lemma. In contrast, if the standardized translation of a term from Arabic was “selfishly”, the suffix “-ly” would be dropped, producing the word “selfish”. Although the suffix “-ish” represents a derivational morpheme, its removal would produce the word “self”, the semantic scope of which no longer captures the mentalistic gloss of the translation. As such, no further morphological reduction of the translation occurred and the word “selfish” served as its lemma. In this way, lemmas which captured the full range of a root’s inflections were produced for all translations while allowing for the possibility that the semantic scope of some inflections may differ sufficiently from that of the root to warrant a distinct lemma. With this finalized list of lemmas in place, a custom Python script was run to assign a lemmatized English translation to each token in the Raw Data File. With these lemmas, data suited to the analysis of cross-linguistic differences in the words used to describe the video stimuli were generated.

Data Analysis

Poisson variance components models, or pure random-effects models with random intercepts terms for *Participant ID*, *Video ID*, *Field Site*, the interaction between *Video ID* and *Field Site*, and an offset term to control for *Total Length* of transcripts in words were built to ascertain the proportion of total variance in per-transcript LR3PMS counts attributable to these predictors. This offset term was included to account for variability in the total length of the transcripts, a modeling decision which effectively assumed there is some constant rate at which LR3PMS are produced and that the total number of words uttered places a constraint on the maximum count of LR3PMS. To summarize, the modeling approach here is the same as in the previous chapter. Presently, I compared three general linear variance components regression models with Poisson error distributions, or general linear random-effects models with Poisson error distributions, the aim of which was to determine whether there existed differences in the

count of LR3PMS across predictor levels, adjusting for the variable length of each transcript. Additionally, the modal conditional estimates of categorical factor levels were extracted and plotted to determine whether predicted adjustments from the grand mean varied significantly between levels. Next, predicted counts of LR3PMS were obtained using simulated data holding transcript length constant at the lower quartile, median, and upper quartile values of observed transcript length to examine modeled effects under the assumption of no mean differences in total words uttered. Finally, predicted counts of LR3PMS were obtained for each level of *Field Site* by each level of *Video ID* holding *Total Words Uttered* constant at the median transcript length of 85 words. These counts were then rank-ordered to determine whether the extent to which individual videos elicited mental state talk varied across field sites.

Results

Overview of Fitted Random Effects Models

In the present analysis, I compared three general linear variance components models of the frequency with which participants produced LR3PMS. The maximal model contained all four of the following predictor variables: *Participant ID*, *Video ID, Field Site*, and the interaction between *Video ID* and *Field Site*. Across all three models, these variables were treated as random effects with varying intercepts. Models were run in R (Version 4.2.2) using the glmer function from the lme4 package. All models were built with Poisson probability distributions, as the dependent measure was a count variable. Consequently, each model also contained an offset term to account for the fact that transcripts varied in length. To the extent that the production of LR3PMS was a function of the number of exposures (i.e., words uttered), all models required an offset term to control for the total number of words uttered. The candidate models to be compared were all specified according to theory-driven concerns and their ability to speak to questions about the relative importance of individual-level, speech-subject-level, and

cultural-linguistic-context level drivers of variation in LR3PMS. The models were compared and selected according to the results of a Chi-squared goodness of fit test.

Variance Component Model Comparison

A total of three separate variance component models, or pure random effects models, were run to determine the proportion of the total variance accounted for by each predictor in the model, as well as to derive the estimated variance in each predictor. Three models were run to determine the effect the predictors had on each other with respect to the estimated variance attributable to each factor. Those predictors for which the effects were broadly consistent across models were interpreted as being stronger drivers of variance in participants' production of LR3PMS than others. Those predictors for which the effects changed according to the inclusion of other factors were interpreted as being less meaningful determinants of when participants produced LR3PMS. Models were generated according to their ability to illustrate what component of the variance was explained by *Field Site* and *Video ID*, conditional on the other factors included in the model. All models were run as Poisson regression models and using the 'glmer' function of the 'lme4' package (Version 1.1-35.3) in the R statistical programming language (Version 4.4.1). The models, in order, were specified as follows:

1. *LR3PMS ~ (1|Video ID*Field Site) + (1|Video ID) + (1|Field Site) + (1|Field Site / Participant ID) + Offset(Log(Total Words Uttered))*
2. *LR3PMS ~ (1|Video ID*Field Site) + (1|Video ID) + (1|Field Site) + (1|Participant ID) + Offset(Log(Total Words Uttered))*
3. *LR3PMS ~ (1|Video ID) + (1|Field Site) + (1|Participant ID) + Offset(Log(Total Words Uttered))*

The terms in these models indicate that the count of LR3PMS was modeled as a function of random intercepts for each level of *Video ID* (specified as *(1|Video ID)*), random intercepts for each level of *Field Site* (specified as *(1|Field Site)*), random intercepts for each

level of the interaction between *Video ID* and *Field Site* specified as *(1|Video ID*Field Site),* and random intercepts for each level of *Participant ID* (specified as *(1|Participant ID),* each of which was also allowed to vary across *Field Site* to account for the nesting structure between these two variables (specified as *(1|Field Site/Participant ID)* though equivalent to *(1|Participant ID),* as will be illustrated in the following two sections). Additionally, an offset term was included to account for the fact that the greater the number of words uttered, the greater the number of exposures within which a LR3PMS might or might not be logged. In effect, the offset term accounted for the intuition that a transcript containing ten LR3PMS out of one hundred total words uttered might differ meaningfully from a transcript containing ten LR3PMS out of one thousand total words. *Participant ID* served as a predictor in all three models as the transcript data constituted repeated measures.

Variance Component Model 1 (VCM 1)

As the full model against which the remaining two would be compared, VCM 1 was fit to examine the role of *Participant ID, Video ID, Field Site,* and the interaction of *Video ID* and *Field Site* simultaneously. VCM 1 fit was evaluated using the AIC, or Akaike Information Criterion (AIC = 6143.2) and the BIC, or Bayesian Information Criterion (BIC = 6170.0). The log-likelihood of the model was also reported (log-likelihood = -3066.6). Variance estimates and standard deviations for each predictor in the model were assessed to determine the variability captured by the random effects, the results of which are presented in **Table 4**. Variances correspond to the spread in the intercepts across the levels of each predictor. The variance of the random effect may thus be understood as the degree to which the intercepts of each level of a predictor vary. The lower this number, the lower the variation across levels of the predictor.

With this interpretation in mind, the variables for which variation was greatest between levels were *Video ID* (var = 0.0915, sd = 0.3025) and the interaction between *Video ID* and *Field Site* (var = 0.05245, sd = 0.2290), respectively. Crucially, the variance estimates of the

remaining variables exhibited three features of note. First, the variance estimate of *Participant ID* nested within *Field Site* (var = 0.03701, sd = 0.1924) was substantially greater than that of *Field Site* alone (var < 0.0001, sd < 0.0001). Second, both of these values were substantially lower than either of the estimates for *Video ID* or for the interaction between *Video ID* and *Field Site.* Lastly, the variance estimate for *Field Site* alone was effectively zero, a result which was similar to the findings presented in Chapter 3. This state of affairs is the consequence of singular model fit, or a circumstance wherein some dimensions of the variance-covariance matrix were estimated as exactly or very nearly zero. While models with singular fit are statistically well-defined, as it is theoretically sensible for the true maximum likelihood estimate to correspond to a singular fit, such fits may correspond to overfitted models with poor power. This problem is one to which I will return later in the discussion to adjudicate whether it indicates poor model fit and how it ought to be understood in light of model-building considerations.

Intraclass Correlation Coefficients were calculated for each random effect to determine the proportion of the total variance explained by each. 20.45% of the total variance explained by the model was attributable to *Participant ID* nested within *Field Site*, 28.98% was attributable to the interaction between *Field Site* and *Video ID*, 50.56% was attributable to *Video ID*, and effectively none of the variance was attributable to *Field Site*. **Figures 14, 15,** and **16** illustrate the conditional modes of the random intercept estimates with 95% confidence intervals for each of the variables included in the model. The conditional modes correspond to the deviation of a specific group's intercept from the overall average intercept, conditional on the data. In essence, the values presented in **Figures 14 – 16** represent simultaneously the extent to which each level of the predictor differs from the overall average intercept and the extent to which these levels differ from each other.

As can be seen in **Figure 14**, the 95% confidence intervals for all three field sites overlap entirely with each other and are centered on zero. Collectively, these results indicate

that the conditional modal estimates for each level of *Field Site* differ neither from each other nor from the intercept estimated for the overall average. In **Figure 15**, the 95% confidence intervals for each of the nine video stimuli indicate that the only random intercepts reliably different from that of the overall average intercept are those for the Sickness and Norm Violation videos, though the 95% confidence interval for the Prestige and Mate Guarding videos overlap with zero only very slightly. These results mean that the only video stimuli which featured mental state talk at a frequency significantly different from the grand mean were the sickness and Norm Violation videos. Descriptions of the Sickness video had more mental state talk than the overall average while descriptions of the Norm Violation video had much less. The confidence intervals on the conditional modal intercept estimate for the Mate Guarding, Prestige, and Sickness videos overlap with those of every other video, while the confidence intervals on the conditional modal intercept for the Norm Violation video is reliably different from every video other than Infidelity and Dominance. These results suggest that at least one of the video stimuli tended to elicit fewer LR3PMS than the overall average and at least one of the video stimuli tended to elicit more LR3PMS than the overall average.

In **Figure 16A,** it can be seen that none of the conditional modal estimates of the random intercepts for *Video ID* reliably differed from zero or from each other among participants recruited from China. The same can generally be said for participants recruited from the United States, with the exception of the conditional modal estimate of the intercept for the False Belief video stimulus which was found to be reliably below average and the Prestige video stimulus which was found to be reliably above average. In contrast, four of the conditional modal estimates of the random intercepts for *Video ID* differed reliably or nearly reliably from zero among participants recruited from Morocco. Among these participants, estimates were higher than the overall average intercept estimate for the False Belief video stimulus and nearly reliably higher for the Mate Guarding video stimulus, while estimates were lower than the overall

average intercept estimate for both the Norm Violation and Dominance video stimuli. **Figure 16B** presents the same data as **Figure 16A** grouped by *Video ID* on the y axis and helps to illustrate that across all 9 video stimuli, the conditional modal estimates of the random intercepts for each country do not reliably differ from each other across any of the video stimuli with the exception of the Dominance video stimulus.

Variance Component Model 2 (VCM 2)

Given the results of VCM1 which suggested that only a minimal fraction of the variability in the data was attributable to *Field Site*, VCM 2 was built to determine whether the structure of the data wherein *Participant ID* was nested within *Field Site* resulted in the misattribution of variability in *Field Site* to *Participant ID.* As such, VCM 2 sought to replicate the overall structure of VCM 1 while treating *Participant ID* as non-nested. The results of VCM 2 were identical to those of VCM 1. The resulting table and figures can be found in Appendix C.

Variance Component Model 3 (VCM 3)

A primary finding of VCM 1 was that nearly 80% of the variance in the data was attributable to the combined effect of *Video ID* and the interaction between *Video ID* and *Field Site*. Critically, none of the variance in the data was attributed to *Field Site* alone. While the result of VCM 1 provided some evidence to suggest that *Field Site* is a relatively unimportant predictor of variance in the count of LR3PMS within transcripts, it was possible that the random effect term for *Field Site* and the random effect term for the interaction between *Video ID* and *Field Site* were in fact highly collinear and thus competing to explain the same variance. If so, then removal of the interaction term from VCM 1 ought to result in the reapportionment of its attributed variance to *Field Site*. VCM 3 sought to determine whether this was the case. VCM 3 fit was evaluated using the AIC, or Akaike Information Criterion (AIC = 6265.7) and the BIC, or Bayesian Information Criterion (BIC = 6287.1). The log-likelihood of the model was also reported (log-likelihood = -3128.8). Variance estimates and standard deviations for each

predictor in the model were assessed to determine the variability the random effects captured, the results of which can be found in **Table 5**.

With the exclusion of the random effect term for the interaction between *Video ID* and *Field Site*, the ordering of the variables by the amount of variance attributed to them was the same for VCM 3 as it was for VCM 1. The variance estimates were greatest for *Video ID* (var = 0.0967, sd = 0.3110), followed by *Participant ID* (var = 0.0373, sd = 0.1932) and finally by *Field Site* (var = 0.0024, sd = 0.0486). Perhaps unsurprisingly, the amount of variance attributed to *Field Site* did increase relative to VCM 1, although the increase did not represent a simple transfer from the interaction term to the *Field Site* term, but a complex reapportionment in which a substantial amount of explained variance was lost. Thus, while it seems to be the case that some of the variance attributable to the interaction term overlapped with that of *Field Site,* a substantial proportion was uniquely attributable to the particular effects of particular videos within each field site. Nevertheless, VCM 3 reaffirmed the finding of VCM 1 that *Video ID* appears to account for a greater degree of variance in the data than does *Field Site.* Intraclass Correlation Coefficients were calculated for each random effect to determine the proportion of the total variance explained by each. 27.35% of the total variance explained by the model was attributable to *Participant ID*, 70.92% was attributable to *Video ID*, and 1.73% was attributable to *Field Site*. While this represents a notable increase from the proportion attributable to *Field Site* in VCM 1, it remains accountable for substantially less variance than *Participant ID* or *Video ID.* **Figures 17** and **18** illustrate the conditional modes of the random intercepts estimates with 95% confidence intervals for each of the variables included in the model, untransformed. The conditional modes correspond to the deviation of a specific group's intercept from the overall average intercept, conditional on the data. In essence, then, the values presented in **Figures 17** and **18** represent simultaneously the extent to which each level of the variable differs from the overall average intercept and the extent to which these levels differ from each other.

As can be seen in **Figure 17**, the 95% confidence intervals for the conditional modal estimates of the intercept for all three of the field sites overlap substantially with zero and with each other. Collectively, these results indicate that the conditional modal intercept estimates for the United States, China, and Morocco do not differ reliably from the overall average intercept nor do they differ reliably from each other. In **Figure 18**, the 95% confidence intervals for each of the nine video stimuli indicate that the conditional modal intercept estimates of only one video stimulus overlaps substantially with zero. Thus, eight videos have conditional modal intercept estimates that reliably differ from that of the overall average intercept. These videos, in order of the absolute magnitude of difference from the overall intercept, are the Norm Violation video stimulus (lower than average), the Prestige video stimulus (greater than average), the Sickness video stimulus (greater than average), the Mate Guarding video stimulus (greater than average), the Infidelity video stimulus (lower than average), the Dangerous Animal video stimulus (greater than average), the Dominance video stimulus (lower than average), and the False Belief video stimulus (greater than average). Of these conditional modal intercept estimates, those of the Sickness, Prestige, Mate Guarding, False Belief and Dangerous Animal video stimuli did not reliably differ from each other, though they did differ from the Norm Violation, Infidelity, and Dominance video stimuli. The conditional modal intercept estimates for the Infidelity and Dominance video stimuli did not reliably differ from each other, though they did from the Norm Violation video stimulus. In total, these findings point toward meaningful variability in the extent to which the video stimuli tend to elicit LR3PMS.

Based on both the fit statistics and the interpretation provided above, VCM 1 represented a better fit to the data, with lower values than VCM 3 across AIC, BIC, and log likelihood scores. Moreover, VCM 1 accounted for a greater proportion of the variance in the count of LR3PMS within transcripts than VCM 3, the results of which were confirmed with a Chi-Square difference test that was highly statistically significant, $X2\ (1, N = 2) = 124.48$, $p < .0001$.

This result indicates that the larger model (VCM 1), with a greater number of estimated parameters fits the data more closely than the smaller model (VCM 3). Therefore, the inclusion of random intercepts for each level of the interaction between *Video ID* and *Field Site* accounted for a sufficient amount of variation in the count of LR3PMS, over and above that of random effects for just *Field Site* and *Video ID* alone, and therefore motivated proceeding with VCM 1 in subsequent analysis of the data. A consequence of this finding was that the best fit model, VCM 1, was one in which the amount of variance attributable to *Field Site* independent of the effect of *Video ID* was, very nearly zero. This result will be discussed in greater detail in the discussion.

Evaluation of Model Fit

Figures 19 and 20 illustrate both the observed and fitted mean counts of LR3PMS for the interaction between *Video ID* and *Field Site* (**Figure 19**), for the main effect of *Video ID* (**Figures 20C and 20A**), and for the main effect of *Field Site* (**Figures 20D and 20B**). Notably, the fitted estimates of the mean count of LR3PMS among transcripts corresponding to each of the levels of these predictors were very close to the observed values, further indicating good model fit. Next, a simulated dataset was generated to ascertain the predicted average count of LR3PMS for *Video ID*, *Field Site*, and their interaction when transcript length was held constant. Transcript length values corresponded to the lower quartile (n=40 words), the median (n = 85 words), and the upper quartile (n = 142 words) of observed transcript lengths. The simulated dataset contained 4779 observations corresponding to three transcripts varying in overall length (40 words, 85 words, and 142 words) per participant (n=177) per video stimulus (n = 9). Field site was left unmanipulated across participants to account for the fact that they could not have been drawn, counterfactually, from different sites, though it is plausible to imagine that they might otherwise have spoken to greater or lesser extents than actually observed.

The simulated dataset was fed into VCM 1 and the resulting predictions, with standard errors, were used to produce mean predicted counts of LR3PMS and associated standard

errors for each level of the predictors in the model. These results were plotted and can be found in **Figures 21, 22, and 23**. Predicted counts of LR3PMS for the interaction between *Field Site* and *Video ID* across transcripts at the lower quartile value, the median value, and the upper quartile value of transcript length can be found in **Figures 21A and 21B.** As can be seen most clearly in **Figure 21A**, the confidence intervals for almost every level of the interaction term, at each of the three specified transcript lengths do not include zero, indicating that predicted counts of LR3PMS are reliably greater than zero. Notably, however, the vast majority of the levels of the interaction term do not reliably differ from each other, with the except of the Norm Violation video stimulus across all field sites at all transcript lengths. **Figure 21B** presents the same data organized by *Field Site* on the x axis.

Next, observations were collapsed across *Field Site* to view the predicted count of LR3PMS for each level of *Video ID* at transcripts lengths of 40, 85, and 142 words. As can be seen in **Figure 22**, the predicted count of LR3PMS was reliably different from zero for all levels of *Video ID* across all transcript lengths. The predicted count of LR3PMS for the Dangerous Animal, False Belief, Mate Guarding, Prestige, and Sickness video stimuli did not reliably differ from each other. Similarly, the predicted counts for the Dominance and Infidelity video stimuli did not reliably differ from each other. The predicted count of LR3PMS for the Norm Violation video stimulus reliably differed from both of these clusters. Across transcript lengths, the Norm Violation video stimulus yielded predicted counts of LR3PMS reliably lower than all other video stimuli. In **Figure 23**, observations were collapsed across *Video ID* to quantify the predicted count of LR3PMS for each level of *Field Site* across transcripts of 40, 85, and 142 words in length. Across all transcript lengths, the average predicted count of LR3PMS in transcripts produced by participants from China, Morocco, and the United States were reliably above zero but not reliably different from each other. Surprisingly, mean predicted counts of LR3PMS were higher for China when holding transcript length constant at the first quartile, median, and third

quartile values. This finding represented a departure from the observed (**Figure 20B**) and fitted values (**Figure 20D**) wherein the mean count of LR3PMS was lower for transcripts produced by Chinese participants when compared to those produced by American participants and higher when compared to those produced by Moroccan participants.

To shed light on this contradictory finding, mean transcript lengths were plotted as a function of *Video ID*, *Field Site,* and the interaction between the two variables. Description of the observed data in this way is presented in **Figures 12 and 13**. A striking difference in the average length of transcripts produced by participants in Morocco relative to participants in China or the United States is especially readily observed in **Figures 12A and 13B.** Averaging across videos, transcripts produced by Moroccan participants are reliably shorter than those produced by American Participants and nearly reliably shorter than those produced by Chinese participants (**Figure 13B**). This same pattern holds at least as strongly, if not more so, when broken down by the particular video stimuli to which a transcript corresponds (**Figure 12A**). Additionally, **Figures 12B and 13A** appear to suggest that the rank ordering of video stimuli by mean transcript length is more or less the same across the levels of *Field Site*, indicating that the total number of words uttered may be tracking a property of the video stimuli themselves, such as duration in seconds. Pearson's product-moment correlation was conducted on these data and a small, but highly statistically significant relationship was found between the total number of words uttered in transcripts and the length in seconds of the video to which the transcript corresponded, $r(1587) = .122$, $p < .0001$. Thus, though the mean length of transcripts varied substantially across field sites, these values may have been tracking structural features of the content to which they corresponded – in particular, the length of the video stimuli. However, subsequent correlation analyses suggested that the total number of words uttered within a given transcript was far more strongly correlated with the count of LR3PMS, $r(1587)=0.755$, $p<.0001$.

Finally, I sought to test whether some of the video stimuli elicited mental state talk more strongly than others across field sites. Given the greatest amount of variance in VCM 1 was attributed to the *Video ID* predictor, it is unlikely to be the case that there would be a significant reordering of the video stimuli with respect to the count of LR3PMS they elicited. To test this assumption, predicted counts of LR3PMS were obtained from VCM 1 for each level of *Field Site* by each level of *Video ID* holding *Total Words Uttered* constant at the median transcript length of 85 words. Within each level of *Field Site*, the levels of *Video ID* were rank-ordered according to predicted count of LR3PMS and correlation analyses were conducted on rank orderings between each level of *Field Site*. These tests allowed me to determine the extent to which individual videos elicited LR3PMS varied across field sites. Three correlation analyses were run, comparing the rank ordering of video stimuli by predicted count of LR3PMS between China and Morocco (rho = 0.933), the United States and Morocco (rho = 0.45), and the United States and China (0.567). These results were further confirmed by examining scatterplots of the rank ordering of the stimuli for each of these pairings of field sites, which seemed to suggest that the rank ordering of the video stimuli was relatively stable across field sites (**Figure 24**).

Discussion

Overall, the analyses presented above provide preliminary evidence for four claims. The first is that although the absolute counts of LR3PMS vary across field sites, the *rate* at which mental state talk occurs appears to be yoked to the total amount of speech independent of cultural-linguistic context when coding for all mental state terms. Understood as such, the frequency of mental state terms might be understood as occurring at a more or less fixed rate across the languages and field sites sampled. The second is that individuals vary in how often they talk about mental states. The third is that the amount of mental state talk participants produce is most strongly predicted by the video stimulus they are describing. The fourth is that the extent to which a given video stimulus elicits mental state talk varies across the field sites

from which Moroccan Arabic, American English, and Mandarin Chinese speakers were recruited. One possible interpretation of this finding was that even though the actual frequency of mental state talk for a given video varied across the three field sites, the amount of mental state talk the video elicited relative to the other videos was the same across sites. In effect, it was possible that the rank ordering of the video stimuli was the same within each field site, but the extent to which a given video stimulus elicited LR3PMS varied across field sites. This hypothesis was tentatively confirmed through correlation tests which showed broad-cross-cultural similarity in the video that elicited the least mental state talk, the video that elicited the 2nd-least mental state talk, the 3rd-least mental state talk, and so on. The magnitudes of the three correlations were relatively large, though only one of the three tests performed achieved statistical significance. Thus, my findings also provide evidence for a fifth claim. Namely, the content of the video stimulus strongly structures whether or not mental state talk will occur, though the cultural-linguistic environment appears to mediate the extent to which a given video stimulus will elicit such talk.

Implications of My Findings for Outstanding Questions in the Extant Literature

The finding that *Participant ID* explains a meaningful degree of variance in the production of LR3PMS is consistent with the notion that there exists significant variation in how much people talk about the mind within populations, when measured according to the video elicitation task employed here. That the frequency of speech about the mind within populations may exist as a distribution across individuals parallels claims that have been made about mindreading more generally. Namely, that there exists a distribution of mindreading phenotypes within populations. Though it may appear obvious that the frequency of mental state talk would vary across individuals, there were theoretical reasons not to treat this assumption as granted. Given the role of mindreading in facilitating the capacity for language (Kwisthout et al., 2008; Scott-Phillips, 2010, 2014; Scott-Phillips et al., 2009; Seyfarth & Cheney, 2014), as well as its

centrality to many of the phenomena argued to be crucial to the fitness of human beings (Barrett et al., 2010; Caballero et al., 2013; Carpenter & Tomasello, 1995; Csibra & Gergely, 2006; Paal & Bereczkei, 2007; Southgate et al., 2009), it was plausible that whatever variation in mental state talk exists might be below the threshold of detection. As *Participant ID* was found to account for a significant proportion of variance in the production of LR3PMS, this finding is consistent with earlier results which have suggested that behavioral differences between individuals represent an important axis along which variation in mental state talk manifests (Bretherton & Beeghly, 1982; Hughes & Dunn, 1998; Lecce et al., 2021; Pennebaker & King, 1999; Ruffman et al., 2002). This finding strengthens the conclusion that the production of LR3PMS may be similar to other aspects of the mindreading capacity which have previously been shown to vary between individuals (Baron-Cohen et al., 2003; Baron-Cohen & Wheelwright, 2004; Taylor & Carlson, 1997; Turner & Felisberti, 2017; Woo et al., 2023). More narrowly, the fact I observed individual differences in the production of LR3PMS is in some ways a replication of the claim there exists variation in the frequency with which adults produce lexical references to others' mental states, as reported by Ruffman et al., (2002). My replication of this finding suggests the soundness of at least one of the presumptions these authors made.

Furthermore, the finding that *Video ID* accounts for a significant proportion of variance in the count of LR3PMS is consistent with claims that have been made which purport certain subjects or topics of speech recruit the same concepts, categories, or cognitive capacities in relatively invariant ways across cultural contexts (Floyd et al., 2018; Goddard, 2010; Huang & Jaszczolt, 2018; Imai & Gentner, 1997; Jackson et al., 2019). Research in psychology suggests that situational context can have a major impact on the production of certain word types, such as function words (Frank et al., 2013; W. S. Hall et al., 1981; Hawkins & Goodman, 2016; Y.-S. G. Kim et al., 2021; A. Lindström & Sorjonen, 2012; Parrigon et al., 2017; Pennebaker et al., 2003; Roby & Scott, 2022; Stivers et al., 2011). But findings on whether this holds true for

content words have been limited. Nevertheless, there are straightforward reasons to think this phenomenon may generalize. For example, it seems clear that conversations occurring in the context of jury trials on defendant culpability might more strongly elicit terms referring to mental states than conversations occurring in the context of a quarterly earnings report (Baetens et al., 2014; Conley, 2015; Vásquez & Urzúa, 2009). These may represent contrived examples, but they provide a framework from which to begin an investigation of speech about naturalistic, everyday interactions. Though there is some ethnographic research to suggest the importance of speech content on the production of mental state talk (Conley, 2015), these findings are among the first to suggest experimentally that the frequency of such LR3PMS is indeed predominantly a function of speech content, or the situations about which speech is generated, and not of individual differences between participants or differences between cultural-linguistic contexts.

My findings that *Field Site* has no main effect on the count of LR3PMS and that the interaction between *Video ID* and *Field Site* accounts for more variation than *Participant ID* but less than *Field Site* alone presents a complicated picture with respect to the extant literature. That there is no main effect of Field Site appears to suggest that cultural-linguistic context accounts for little or no variation in the production of LR3PMS. This finding, when interpreted in isolation of my other findings, runs counter to some of the claims that have been made in the literature about specific ways in which cross-linguistic and cross-cultural variation in mental state talk might manifest (Cheung et al., 2009; Jackson et al., 2019; Schwanenflugel et al., 1994). While it remains to be determined precisely how universal emotional and mental state concepts and categories are, my data cannot speak to these particular debates (Floyd et al., 2018; Goddard, 2010). However, my findings can illuminate further some of the debates that have made explicit claims about variation in the overall quantity of mental state talk.

Given that my data showed absolute differences in both overall count of LR3PMS and overall number of words when comparing between field sites, it may well be the case that the ethnographic data is accurate with respect to such absolute differences in the quantity of LR3PMS. As claims made by ethnographic researchers are also a function of the phenomena to which they attend, it may also be possible that the relative significance of a phenomenon is over- or understated. For example, it is possible that the raw number of such attributions may indeed be lower in one language community than in another. If, however, the ethnographer fails to attend to the fact that members of the community with whom they work are also less talkative in general, then the nature of their attention has obfuscated the possibility that there may be no differences across language communities with respect to the rate or relative frequency of mental state attributions. The existence of variation in the pragmatically structured communicative norms across language communities is well-documented, and it is possible that the overall quantity of information provided in speech about others' mental states is free to vary while there remain more universal constraints on the requisite quality of information to be conveyed about others' mental states. The current results are equivocal with respect to how well they resolve the uncertainty surrounding such earlier ethnographic claims. While my findings did not demonstrate a main effect of *Field Site* with respect to the count of LR3PMS when controlling for total transcript length, the same cannot be said if this variable is left out of the model. When excluding an offset term for description length, I found a significant difference between the United States and Morocco in the count of LR3PMS, indicating that without consideration of the overall verbosity of participants from a given field site, there may indeed emerge differences suggestive of the kinds of findings reported in the ethnographic data.

However, the fact that 1) a significant proportion of variance was attributed to the interaction between *Field Site* and *Video ID* and that 2) the rank-ordering of video stimuli by their predicted counts of LR3PMS in China, Morocco, and the United States were all strongly

correlated with each other appears to suggest that the relative extent to which a given video stimulus elicits LR3PMS (with respect to the other video stimuli) is similar across field sites. The absolute extent to which a given video stimulus elicited LR3PMS across the field sites did vary, as can be seen in both the observed and fitted values of the data (**Figures 38, 39, 40, and 41**). Thus, the field sites from which I drew my sample varied in the absolute counts of LR3PMS, though this finding raised additional questions as to why that was the case.

An assessment of transcript length demonstrated that there are systematic and substantial differences in the mean number of words produced by participants from each field site (**Figures 50, 51, and 53**). Additionally, it was found that the count of LR3PMS was strongly correlated with the total number of words uttered. Consequently, one potential reason there exist absolute differences in the count of LR3PMS is because there exist absolute differences in the volume of speech produced. This result itself raises further questions, the answers to which may be as simple as variation across field sites in participant comfort with the research setting (L. Milligan, 2016) or may be of more substantial empirical and theoretical interest. One possibility, which is itself related to participant comfort, may be that participants across field sites are bringing to bear distinct goals of memory retrieval when providing their narrative descriptions of the video stimuli. As demonstrated by Dutemple and Sheldon (2022), encoding of the stimulus with a social goal as opposed to a goal of accuracy in subsequent retelling results in the exclusion and reordering of details in the original stimulus. People in different societies might vary in how comfortable they are in the experimental setting. This could impact the naturalness of their speech, and could lead to differences in length or content. Participants' ease during testing might also impact the way their attention is directed during the video viewing task. Previous research has illustrated cross-cultural variation in attention to visual stimuli when tasked with narrative construction of the observed stimuli (Cohn et al., 2012; Senzaki et al., 2014). As such, it is possible that the differences in narrative length reflected differences in the

allocation of attention to various details in the video stimuli and their subsequent elision or inclusion in the narrative retellings. Crucially, preliminary qualitative analyses of participant descriptions show no notable differences in macrostructural features or inclusion of the central narrative elements, consistent with earlier findings (Chang, 2009; Gorman et al., 2011; Méndez et al., 2023; J. G. Miller, 1986).

Additionally, my findings allowed me to rank predictors by the amount of variance they each explained, giving me indirect evidence of their relative importance in driving the production of LR3PMS. I found that *Video ID* explained the greatest amount of variation, the interaction between *Video ID* and *Field Site* explained the second greatest amount of variation*, Participant ID* explained the next greatest amount of variation, and *Field Site* alone explained effectively none. These findings suggest that what one talks about is the single greatest determinant of how frequently LR3PMS occur in elicited narrative descriptions of my video stimuli. The cultural-linguistic context within which an individual is situated appears to shape the absolute amount one says about mental states for a given topic, and this topic-level impact appears to drive more variation in the production of LR3PMS than individual-level variation in the propensity to produce LR3PMS. In effect, then, I present evidence that the cultural-linguistic contexts from which my participants were drawn do not, in and of themselves, drive variation in the frequency with which LR3PMS are produced. However, the cultural-linguistic context may drive variation in ways that are conditioned on the specific topic or subject of speech such that the count of LR3PMS for a single topic or speech subject may vary substantially across field sites. Nevertheless, the relative extent to which a given topic or subject elicits LR3PMS when compared to another seems well-preserved across cultural-linguistic contexts. This is reflected in the variation attributed to topic or subject of speech (as indexed by Video ID), which was the greatest of all the predictors in my model. Finally, a fair degree of variation in the production of LR3PM between individuals was found.

This picture is consistent with a number of broader theoretical claims that have been made in anthropology, sociology, and psychology. My findings can be understood through the lens of interactionism, which posits that social behavior is an interactive product of individuals and situations (Berge & Raad, 2001; Murtha et al., 1996; Sherman et al., 2015). Under this view, human behavior can be partitioned into three parts: Traits, or the extent to which properties of the individual, like personality, directly affect behavior; situations, or the extent to which any given person will provide basically the same response to a given situation; and interactions, or the way in which the same situation affects individual people differently. Under this interpretive framework, the situation as defined by the narrative retelling of a particular video stimulus accounts for the greatest variation in the behavior of producing LR3PMS. If one's individual propensity to produce LR3PMS is understood as a trait, then the extent to which it determines the behavior of producing LR3PMS is dwarfed by the situation itself. However, an individual's membership in a particular culture, if understood as a trait, interacts with the situation such that membership in one culture or another serves to condition the absolute volume of LR3PMS in variable, situation-dependent ways.

The fact that the rank-ordering of the stimuli across data from the three field sites is broadly similar also coheres with claims that have been made in cultural attractor theory. Namely, my data appear consistent with the notion that individuals across cultural-linguistic contexts are equipped with a shared, universal cognitive apparatus to make sense of social interactions, the outputs of which are biased in the extent to which observed interactions between others are glossed through the lens of their mental states, though not deterministically so. Thus, the relative quantity of LR3PMS used to describe a given naturalistic social interaction when compared to another may be largely the same across cultural-linguistic contexts. However, the absolute quantity may vary across cultures in ways that are caused by or linked to other cultural phenomena (Barron & Schneider, 2009; Hansen et al., n.d.; Levinson et al., 1987;

Mehl et al., 2007; Newman et al., 2008; Robbins, 2004). In summary, my results provide preliminary evidence to suggest that when controlling for volume of speech, the frequency of LR3PMS does not differ appreciably across the set of field sites/languages sampled here.

Implications of My Findings for Those Reported in Chapter 3

Broadly speaking, the results of these analyses are consistent with those reported in Chapter 3. Regardless of whether the data is coded using the eight-word coding scheme of Chapter 3 or the more inclusive scheme employed here, the variance in the count of LR3PMS attributed to *Field Site* alone is minimal, the variance attributed to *Participant ID* is intermediate, and the vast majority of the variance is attributed to *Video ID* and the interaction between *Video ID* and *Field Site*. Where the present results diverge from those of Chapter 3 is in the observed and predicted counts of LR3PMS in each of the field sites. Using the narrow coding scheme inspired by Wellman and Estes (1987), no mean differences in the count of LR3PMS were observed across the Moroccan, Chinese, and American samples. When predictions based on simulated data where transcript length was held constant, however, the predicted count of LR3PMS for Moroccan participants was significantly higher than the predicted counts for Chinese or American participants. In contrast, substantial mean differences in the observed count of LR3PMS were observed across the Moroccan, Chinese, and American samples when using the more inclusive coding scheme, though all such values were higher than when using the narrow coding scheme from Chapter 3. However, these differences were attenuated when predicted counts of LR3PMS using simulated data with transcript length held constant. Moreover, the correlation between count of LR3PMS per transcript and number of words uttered per transcript was nearly three times as strong as that observed when coding the data using the narrow scheme from Chapter 3.

These findings, collectively, suggest that when allowing coders who are native speakers of the target languages to identify all instances of LR3PMS, the absolute counts of LR3PMS

vary across field sites but in a way that is yoked to the total amount of speech. Understood as such, the frequency of mental state terms might be understood as occurring at a more or less fixed rate across cultural-linguistic contexts. Notably, the terms included in the narrow scheme were predominantly of a cognitive or epistemic nature, whereas no such restrictions were placed on the terms in the presently employed coding scheme. Consequently, the findings reported here provide preliminary evidence that the specific variety of mental states to which appeals are made in narrative descriptions may vary across cultures such that relatively fewer cognitive terms were used among English and Mandarin speakers when compared to Arabic speakers. When a wider net is cast, the differences in overall counts of mental state terms are attenuated. This interpretation of the data is consistent with claims that have been made about the presence of shared and variable conceptions of mental states (Goddard, 2010; Goddard & Wierzbicka, 1994; Jackson et al., 2019; Wierzbicka, 1992, 1996). These results underlie the importance of using coders who are fluent, first-language speakers of the target languages for tasks of this nature. By leveraging their emic knowledge of the language, phenomena like non-mentalistic metaphors which are, in fact, about the minds of others, but which may not be understood as such etically, this approach increases the likelihood of capturing such references. By using elicited descriptions, I ensure comparability and uniformity in “access” to the set of circumstances depicted in the video stimuli.

Limitations

Initial processing of transcripts

Several of the assumptions that motivated the decision to code lexical items, as opposed to morphemes or phrasal structures, were rooted in an English-biased perspective that took for granted clear word boundaries built into the orthography of the transcripts. This presented problems when dealing with transcripts from each of the other languages sampled. In the case of Mandarin, the standard orthography does not demarcate word boundaries with spaces. As

such, readers of Mandarin needed to determine as they read from context where one word ended and the next began. Consequently, in order to generate a comparable frequency dictionary, coders were asked to draw word boundaries in the transcripts by placing spaces between the start of one word and the beginning of the next. While the coders were provided with guidelines for determining word boundaries, the very fact that such decisions had to be made were potential sources of both systematic and random error in the Mandarin data set. In contrast, while the Arabic data cannot be said to have the same orthographic issues, the increased morphological complexity relative to that of the English data meant the total number of items to be coded in the frequency dictionary was higher, therefore increasing the risk of errors. To further illustrate how this strategy limits my findings, we may also consider the fact that there may be crucial mental state terms that are only ever made manifest in the language in the form of compound words. As such, the strategy employed may systematically miss some such references to the extent that this pattern characterizes the languages sampled.

Another limitation to my findings is that the video stimuli I employed were not not psychometrically calibrated to guarantee comparability of the stimuli with respect to number of characters depicted, situational complexity, and length. As such, they need to be interpreted with caution – some of the stimuli, like the Norm Violation video in particular, may have been less efficacious at getting participants to view the action in a narrative framework. Because this particular video appeared to depict a ceremonial procedure with rote steps, the narrative thread may have been more subtle and thus less readily picked up on by participants. It is plausible that some of the video stimuli were just less readily interpreted by participants regardless of linguistic or cultural background. For these reasons, I cannot conclude definitely that the content of the video is itself driving the variation in mental state talk. Nevertheless, these videos appear to act similarly on speakers across the languages sampled. Further investigations are required to determine the source of this variance.

Conclusion

The present study provides evidence that when controlling for speech length, variance across languages in the frequency of LR3PMS is effectively non-existent and is dwarfed by the variance accounted for by individual differences and the video stimuli being described. The substantial variance attributed to *Video ID*, and the consistency of the rank-ordering of the video stimuli with respect to their predicted counts of LR3PMS within each field site suggests that the content of talk is an important determinant of when people use mental state language, independent of culture or language. Nevertheless, these results also suggest that culture or language can influence how much certain topics of speech elicit third-party mental state talk. While I cannot point to a definitive mechanism by which this effect is achieved, my results provide an initial step towards resolving questions about the way in which mental state talk does and does not vary across language communities. These results suggest that when speakers of Moroccan Arabic, American English, and Mandarin Chinese are asked to describe the scenarios depicted in the video stimuli I developed, they do not differ in the frequency with which they talk about mental states. Interestingly, there exist substantial individual differences in the frequency of mental state talk within each sample of speakers. Given the data generated by western psychologists suggesting a relationship between the production of mental state talk and mindreading ability (Carr et al., 2018; Ruffman et al., 2002), these findings suggest at least one part of this correlations holds true in non-western contexts – namely, while there are no cross-linguistic differences in the frequency of mental state talk when comparing groups of Moroccan Arabic, American English, and Mandarin Chinese speakers, *individuals* differ in the amount of mental state talk they produce. Does this variation correlate with mindreading ability in the same way among these three samples of speakers? In the next chapter, this is the question I aim to address.

Chapter 5: Do Properties of Individuals and Their Cultural-Linguistic Contexts Predict Mindreading Ability?

Introduction

It is at present unknown whether individual differences in how frequently people talk about the mental states of others are related to individual differences in their underlying mindreading ability. While a small number of studies pertinent to this question have provided preliminary evidence of a relationship between the frequency of mental state talk and mindreading ability (Carr et al., 2018; Cheung et al., 2004; de Villiers, 2005; de Villiers & Pyers, 2002; Durrleman et al., 2019; K. Milligan et al., 2007; Ruffman et al., 2002), the scope of these findings is limited given 1) their treatment of false belief as equivalent to mindreading more generally, 2) their failure to clearly define what is meant by mental state talk, and 3) their insufficient sampling of languages other than English. The set of constructs that constitute the mindreading capacity, the ways in which mental states and mental state talk have been defined, and the dimensions along which languages vary are all far greater in number than has been captured by these initial studies. As such, it is premature to say with any confidence that the frequency of mental state talk and mindreading ability are related. Nevertheless, this evidence suggests further research is required to resolve these outstanding claims. In this chapter, I aim to address the limitations of these earlier studies, and in so doing, shed light on two important issues. Using a measure of mindreading other than false belief understanding and leveraging two separate definitions of mental state talk to code participant speech, participants were recruited from three linguistically and culturally unrelated field sites in order to 1) determine whether participants' production of lexical references to third-party mental states (LR3PMS) predicts their performance on a broadly-used measure of mindreading ability and 2) determine whether this relationship, if found, holds across cultural-linguistic contexts.

The Role of Representationalism in Mindreading Research and its Focus on False Belief

The study of theory of mind, known otherwise as mindreading or mentalizing, has its origins in and has been influenced by theories articulated first among philosophers of mind. One such theory which has had an especially significant impact on the trajectory of mindreading research is Representationalism, or the representational theory of mind. Representationalism posits that we do not sense and perceive the world external to ourselves as it is objectively. Rather, sensation and perception are mediated by representations, or internal mental models that denote objects in the world (Nelson, 2023). Representationalism gained purchase among early cognitive scientists studying vision, as it provided a parsimonious explanation of visual phenomena like misperceptions and illusions. Under this view, if the features of some object in the visual field are shown to differ from one's perception of those features, then there must be some secondary entity mediating between and accountable for the difference – i.e., a representation. In effect, these differences were understood as the product of imperfect correspondences between the objects themselves and their representations. If representations could correspond imperfectly to incoming sensory data, then it was theorized that perhaps they needed not correspond to any incoming sensory data at all. In this way, representations could stand in for unobservable objects, including those that are out of view, those that do not exist, and those that could not exist (Dennett, 1978; Fodor, 1981, 1992; Sterelny, 1990). Representations could thus account for cognitive processes about abstract objects beyond the realm of sensory perception, chief among which are the mental states of others.

This insight was soon taken up by developmental psychologists interested in the early age at which children engage in pretend play and produce speech referring to the mental states of others (Bretherton & Beeghly, 1982; Leslie, 1987; Shatz et al., 1983). Both of these behaviors were thought, in different ways, to leverage representations of abstract entities. While bouts of pretend play may feature objects and entities that differ from reality, that are not in view, that do

not exist, or that could not exist, talk about the mental states of other people corresponds to objects for which there is no direct evidence. The representation of abstract concepts has been widely believed to mature slowly over early childhood (though see Borghi et al., 2017 for a summary of ongoing debates concerning the mechanisms by which representations of abstract concepts are constructed). As such, the apparent proficiency with which children as young as two years of age talked about the minds of others and engaged in pretend play required a rethinking of conceptual development, as the extant findings suggested that children could not reliably represent the false beliefs of others before the age of 4 (Wimmer & Perner, 1983). How, then, could children talk meaningfully about the representational mental states of others and engage in pretend play if their metarepresentational abilities were as of yet incompletely developed? The test used, the False Belief Task, was later criticized on the basis of its dependence on linguistic competence, its inability to partition out effects attributable to executive function, and its weakness as a tool for measuring individual differences in mindreading ability (P. Bloom & German, 2000). Subsequent refinements of the measure were developed to address these weaknesses, the application of which provided evidence of functional mindreading capacities in both pre-linguistic children and non-human primates (Baillargeon et al., 2010; Krupenye et al., 2016; Onishi & Baillargeon, 2005). As scholarship in this area developed, the texture and function of representational mental states became clearer. The capacity to represent the mental states of one's interlocutors was understood as a kind of mechanism by which a person could predict others' social behavior (A. Clark, 2013; Koster-Hale & Saxe, 2013). By representing the set of representations contained in an interlocutor's mind, an individual can forecast how an interlocutor will think, act, and feel (Premack & Woodruff, 1978). While the earliest scholars of mindreading attended to a wide range of mental state concepts and representations, the field neglected these other phenomena and shifted its focus toward false belief, or belief representations the content of which differs from the state of affairs

to which it corresponds (P. Bloom & German, 2000) for many years before returning again to these other mental state categories.

The Influence of False Belief on Current Understanding of the Relationship Between Language and Mindreading

The attention given to false belief has had a substantial influence on the way in which the connection of mindreading to language has been understood. This influence has shaped the trajectory of research on mindreading and language in two predominant ways; first, by highlighting that there exists a relationship at all and second, by providing an empirical basis upon which to develop hypotheses privileging the connection of cognitive or belief-like verbs to mindreading while leaving other features of language relatively understudied. Shortly after its introduction, the False Belief Test found frequent employment in studies of sociocognitive differences between typically- and atypically-developing children (Baron-Cohen, 1997b; Baron-Cohen et al., 1985; Castelli et al., 2002; Karmiloff-Smith et al., 1995; Rutherford et al., 2002; Senju et al., 2009; White et al., 2009). Building on work by Leslie (1987) in which a common mechanism was proposed to account for both pretend play and the ability to represent mental states, Baron-Cohen, Leslie, and Frith (Baron-Cohen et al., 1985) posited that atypical development in this representational capacity may account for the verbal and non-verbal communication challenges faced by autistic children. Baron-Cohen et al. (1985) found that autistic children did not reliably pass the False Belief Test, though both typically-developing children and children with Down syndrome did. Crucially, the verbal and non-verbal mental age of the sample of autistic children was in fact higher than those of both the typically-developing children and the children with Down syndrome. Thus, this disparity could not be attributed to intellectual disability more broadly. Subsequent research with other atypically developing populations contributed to a body of evidence supporting Baron-Cohen et al.'s initial conclusion. Children with specific language impairment not attributable to other intellectual or developmental

disability were found to have challenges with mindreading (Nilsson & de López, 2016). Children with Williams Syndrome, a condition in which there is intellectual disability but otherwise precocious linguistic competence do not exhibit the same challenges with mindreading (Karmiloff-Smith et al., 1995). Deaf children of hearing parents who received no exposure to sign language during early critical periods exhibit challenges with mindreading despite no other intellectual disability (Schick et al., 2007). Collectively, these studies and others like them constituted a growing body of evidence supporting the existence of a relationship between mindreading and language, though its directionality remained unclear.

Given these findings, some scholars posited that the ability to represent mental states is dependent on the acquisition and mastery of verbs that share the grammatical property of being able to take sentential complements, or clauses embedded as their subjects or objects. These verbs are of a primarily cognitive or belief-like character (e.g., to think, to know, to believe, to dream, to forget, to remember, to suspect, and so forth), they map semantically onto propositional attitudes (or causally efficacious content-bearing internal states) and when followed by the word “that” (operating here as a complementizer, a functional syntactic category containing words that can be used to turn a clause into the subject or object of a sentence) can take whole propositions as their objects (e.g., “the moon is made of cheese” in the sentence “Jason thinks that the moon is made of cheese”). These researchers posited that verbs of this type could scaffold the insight that the contents of an interlocutor’s mind could differ from one’s own or even differ from the objects in the external world to which those contents correspond. For example, my representation of [Jason thinks that {the moon is made of cheese}] can be true even though nested within it is a representation that is false and that I do not hold myself (i.e., {the moon is made of cheese}). These scholars suggested that variation in one’s competence with these verbs and their syntactic properties may account for variation in the ability to represent the mental states of others (de Villiers, 2005; de Villiers & Pyers, 2002; Gleitman,

1990). Though a handful of studies have examined this hypothesis directly (Cheung et al., 2004; de Villiers & Pyers, 2002), the majority have tended to examine other components of language while still demonstrating significant relationships with mindreading.

The inter-connectedness of phonetics, phonology, morphology, syntax, semantics, vocabulary size, pragmatics, mastery of discursive genre, oral-motor skills, hearing ability, and quality of interaction in language development suggests the possibility that atypical development in any of these areas could have knock-on effects that impact mindreading. For example, a relationship has been documented linking performance on measures of theory of mind and competence with aspects of speech which leverage the ability to recognize the intentions of an interlocutor, such as metaphor, simile, and irony (Happé, 1993). Furthermore, a meta-analysis by Milligan, Astington, and Dack (2007) of 104 studies employing a wide range of language and false belief understanding measures demonstrated that general language ability, semantic ability, receptive vocabulary, syntactic ability, and memory for complements all accounted for statistically significant components of the variance in children's performance on false belief measures. However, these effects sizes were not equivalent, with receptive vocabulary explaining the least variance and memory for complements explaining the most variance. While some authors have argued that this variation may be a product of how effectively each measure isolates their targets of measurement from other language abilities, it may also be the case that mental state verbs which can take sentential complements may bear a unique relation to underlying mentalizing ability over and above that of language ability more generally.

There is evidence that variation in children’s exposure to mental state ***terms*** in parental speech predicts the age at which children first pass the False Belief Test (Ruffman et al., 2002). In this study, the *only* factor that mattered was the raw count of mental state terms produced by children’s primary caregivers, independent of overall talkativeness. This finding is consistent with the literature described above, though there remain a number of unanswered questions.

For one, does the effect of caregiver input on children's mentalizing primarily shape the rate at which children develop the skill, or does the absolute volume of input result in differential mindreading development? The ceiling effects and the low degree of granularity intrinsic to the False Belief Test may well mask variation in mindreading ability that cannot otherwise be observed. Secondly, is it exposure to mental state ***terms*** that matter or is it exposure to mental state ***verbs***? The authors do not provide a clear indication of the items they coded and as such, it is hard to say whether the sentential complement account is supported by these findings. Third, what, if anything, do these findings say about the relationship between the *production* of mental state terms and mindreading ability? Are parents who produce more mental state terms more effective mind readers? Do children who pass the False Belief Test earlier go on to produce more mental state terms in their speech across the lifespan? It seems plausible that individual differences in how frequently people produce such terms in their speech may itself be an index of variation in their mindreading ability, but little research addresses this question.

Mindreading is More Than Just Success on the False Belief Task

The preoccupation of mindreading scholarship with false belief has generated a vast body of literature, the results of which have unequivocally enriched our understanding of social cognition within human beings and across species. Though early developmental research provided strong empirical justification to focus on the False Belief Test as a window into the capacity to represent others' minds, later work showed that a narrow focus on this ability alone failed to capture the variation and complexity of the cognitive constructs entailed by mindreading. As such, other abstract mental state representations and their implications for social cognition have until more recently remained relatively underexplored. Perception, desire, motivation, intention, and emotion are just some of the mental state categories that were sidelined in the initial decades of mindreading research, to say nothing of the sensory stimuli to which we attend for the construction of mental representations, like facial expressions, body

language, direction of gaze, and social context (Brooks & Meltzoff, 2005; D'Entremont et al., 1997; Emery & Clayton, 2001; Sonneville et al., 2002; Stewart et al., 2019; Tomasello et al., 2007; Woo et al., 2023). These topics have all garnered increased interest in the past twenty years, and though some early scholars were indeed interested in phenotypic variation in mindreading among neurotypical children and adults, they were limited by the lack of available tools to measure its full breadth. With increased interest in other aspects of mindreading, an arsenal of tools to measure the representation of other kinds of mental states has emerged (Baron-Cohen et al., 1999, 2001; Dziobek et al., 2006; Jolliffe & Baron-Cohen, 1999; Turner & Felisberti, 2017). Among them are the MASC, the Faux Pas Test, the Strange Stories Task, and the Reading the Mind in the Eyes Test (RMET). Each of these measures has offered new avenues through which to understand mindreading. These tools are not without their flaws, of course, and each has been subject to critiques of its reliability and validity (Quesque & Rossetti, 2020; Turner & Felisberti, 2017). Nevertheless, a major affordance of these newer tools is the ability to detect individual variation in the mindreading ability of neurotypical adults. Though limited, these assays represent major methodological advances and have paved the way for researchers to think about mindreading as a set of graded phenomena both within and across human populations (Apperly, 2012; Hughes & Devine, 2015; Lillard, 1998). Given the goals of the present research, a turn toward mental state representations other than false belief provides a point of entry to study the relationship between mindreading and mental state talk.

One fruitful area for thinking about mindreading beyond cognitive or belief-like mental states is emotion recognition. Though the capacity to recognize and represent emotions is almost certainly distinct from the capacity to represent propositional attitudes like belief, there is evidence that variation in these abilities is substantially attributable to variation in shared underlying cognitive capacities (Turner & Felisberti, 2017). Building on a broad literature highlighting the challenges many autistic individuals face with eye contact, Baron-Cohen et. al.

(2001) posited that eyes may be especially rich sources of social information and that accuracy in extracting social information from eyes may be a trait that varies between individuals. As such, Baron-Cohen et. al. (2001) developed a survey that presented participants with images of eyes and required them to identify the internal states depicted therein. This task was thought to assess participants' ability to use information contained within interlocutors' eyes to construct second-order representations of their mental states. To construct such a tool, a set of 36 images of eyes was drawn from advertisements in British print media sources. These images were presented to participants with one word located at each of the image's corners. These four words were comprised of one target word and three "foil" words, or competitor terms to describe the emotional or mental state of the individual depicted in each image, and the location of the target was word was randomly assigned. These terms were generated by Baron-Cohen et al. and underwent subsequent piloting with a panel of 8 judges to ensure that for each item, at least half of the judges selected the target word. The resulting survey, known as the Reading the Mind in the Eyes Test (RMET), was tested with neurotypical adults in both community and university settings. Initial results suggested that adults vary in the facility or accuracy with which they integrate perceptual information about faces into representations of their corresponding mental states. The RMET has been subject to a variety of criticisms since its introduction (Black, 2019; H. Kim et al., 2022). Nevertheless, its widespread use and acceptance in the literature, coupled with its ability to detect individual differences motivate its use in the present study.

Mental State Talk is More Than Just Cognitive or Belief-Like Verbs

As indicated in the previous section, representations of propositional attitudes do not constitute the totality of mental state representations. The mindreading capacity includes the ability to represent the percepts, desires, motivations, intentions, and emotions of others (Apperly, 2008; Bugnyar et al., 2016; Flavell et al., 1981; Gergely et al., 1995; Gray et al., 2007; Hare et al., 2000; Harrigan et al., 2018; Nichols & Stich, 2003; Perner et al., 2003). Accordingly,

there is no reason to presume talk about the minds of others is singularly constituted by the production of verbs that map semantically onto propositional attitudes and can take whole propositions as their objects. Mental state talk ought to be understood as encompassing a wider range of representations, like perceptions, desires, motivations, intentions and emotions. English terms in these categories often belong to grammatical categories other than verbs, to say nothing of their grammatical category membership in other languages. Though the mental state representations to which cognitive or belief-like verbs correspond may be of a uniquely complex character (Leslie & Happé, 1989; S. A. Miller, 2009; Perner et al., 2003; Sullivan et al., 1994), it is ostensibly the case that all mental states correspond to representations of abstract concepts for which there is no direct evidence. If I say “John is happy” because I see him smiling and laughing, whether I can discern the reason for his happiness is irrelevant to the fact that I have constructed a representation of his internal state. Though a representation of this nature may fail to meet the criteria of a propositional attitude, it appears nevertheless to draw on some of the same representational capacities.

How these terms relate to the mindreading ability, then, is a non-trivial question that has been given relatively short shrift in the literature to date. Consequently, there are reasons to be skeptical of extant claims, as it is not obviously the case that primacy should be given to verbs. Though there exist studies showing variation in language phenomena like emotion term category boundaries (Gendron et al., 2014; Jackson et al., 2019), what variation in non-verb terms means for social cognition is relatively unexplored. What relationship, if any, do these other types of words have to the mindreading capacity? Are there specific relationships between words of a given mental state category and mindreading dedicated to that category (e.g., mastery of emotion words bootstrapping emotion understanding)? Or do belief-like verbs exercise a broad effect on mindreading such that mastery of these terms shapes our ability to represent all mental state categories? Do these relationships hold across languages?

Strategy for Addressing Outstanding Questions in Mental State Talk and Mindreading

At present, there are a number of critical questions that remain unanswered and claims that remain unsubstantiated.

1. It is unclear whether the relationship purported to exist between LR3PMS exposure and the age at which children first pass the False Belief Test differentiates between cognitive or belief-like verbs and other types of mental state terms.
2. Relatedly, whether it is the raw number of belief-like verbs that predicts the age at which the child first passes the False Belief Test or the raw number of any and all LR3PMS is unclear.
3. Regardless of the answer to these first two questions, it is also unclear whether the relationship reported by Ruffman et al., (2002) represents a narrow relationship between mental state talk and the capacity to represent belief-like propositional attitudes (as indexed by performance on the False Belief Test) or if the influence extends to the mindreading capacity more generally.
4. Moreover, Ruffman et al., (2002) argued that the impact of LR3PMS was dissociable from the overall quantity of speech. In light of the findings presented in Chapters 3 and 4 of this dissertation, the veracity of this claim appears to depends on which terms are counted. When coding for belief-like verbs, equivalence in the absolute count of LR3PMS across participants recruited from the United States, Morocco, and China was observed. However, when coding for all LR3PMS, cross-cultural variation in the absolute, but not the relative (e.g., scaled by the total number of words uttered) count of LR3PMS was observed. Consequently, if the claims made by Ruffman et al., (2002) are correct, and if their study was to be replicated across these three field sites sampled here, cross-cultural differences in mindreading ability under one coding scheme and cross-cultural uniformity under the other coding scheme would

be predicted. Consequently, whether the overall quantity of speech is unrelated to the count of LR3PMS depends on what counts as a LR3PMS.

5. It is not well-understood whether production of LR3PMS is related to underlying mindreading ability. Though exposure to such terms has been argued to be related, it is not clear whether individuals who are more capable mind readers talk more about the minds of others.

Consequently, studies aimed at addressing these problems need to obtain standardized counts of LR3PS that allow all mental state talk to be accounted for and that allow for differentiation between the production of belief-like mental state verbs and all LR3PMS. Moreover, a measure of mindreading other than the False Belief test needs to be used, as it is does not permit observation of sufficient variation among neurotypical adult populations.

Here I integrate data presented in the previous chapters of this dissertation with participants' performance on the RMET (Baron-Cohen et al., 2001) in order to examine (1) whether there exists a relationship between the production of LR3PMS and performance on this task, (2) whether this relationship holds across the languages sampled. First, I determined whether there were mean differences in RMET performance across field sites when using the original eight-word coding scheme as opposed to one that allowed the set of items coded as correct to differ between each of the field sites. I then selected an RMET scoring methodology with an eye toward theoretical and methodological soundness while also considering the degree of difference in mean RMET scores across sites. Next, per-subject counts of LR3PMS were generated in two ways – one that counted only cognitive or belief-like verbs and one that counted all LR3PMS. Then, I regressed RMET scores as a function of participant LR3PMS counts (both belief-like verbs and all LR3PMS). Model results using standardized and unstandardized RMET scores were compared to determine if they made diverging predictions.

Because they did not, two final models are reported – one for belief-like verbs only and one for all LR3PMS.

Predictions

Based on the extant literature and the data generated in the previous two chapters of this dissertation, there are several empirical gaps that limit the ability to make strong predictions. However, predictions can be made if a set of assumptions hold true. If it is the case that the count of belief-like verbal LR3PMS are all that matter in predicting mindreading ability, then it should be the case that the fit of the model predicting performance on the RMET from belief-like verbal LR3PMS should not improve when including the total number of words uttered by participants. Moreover, the total number of words uttered ought to be a weaker predictor than the count of belief-like verbal LR3PMS. Additionally, if it is true that the relationship between belief-like verbal LR3PMS and False Belief understanding generalizes to other forms of mindreading, and if it is true that the RMET is an equally valid measure across the three field sites sampled, LR3PMS should only significantly predict performance on the RMET when they are coded for belief-like verbs and not when coded for all mental state terms. Alternatively, the former coding of LR3PMS may simply exhibit a stronger relationship than the latter coding of LR3PMS. Furthermore, including the total number of words uttered by participants should not significantly improve model fit regardless of whether LR3PMS are coded for just belief-like verbs or if LR3PMS are coded for all mental state terms. Moreover, total number of words should not significantly predict performance on the RMET. Finally, it was predicted that there would be no variation across field sites in mean performance on the RMET, nor in the strength of the correlation between RMET performance and the count of LR3PMS. The reason for this is because there are no mean differences across field sites in the count of LR3PMS when coding for belief-like verbs, despite mean differences in the total number of words participants uttered.

Methods

Procedure

As indicated in Chapter 1, the experimental procedure employed to generate the corpus of transcripts used for the analyses in both Chapters 1 and 2 featured administration of the RMET to participants as its final step. After participants had completed viewing and describing all nine video stimuli (as described in Chapter 2), they completed the RMET. In the proceeding sections, I detail the underlying logic for using both the original coding scheme developed by Baron-Cohen et al., and a novel coding scheme designed to address issues with the original. I then detail how participant RMET data were processed for subsequent analysis.

Reading the Mind in the Eyes Test

RMET data were collected from participants as part of a broader interview, the details of which can be found in Chapter 2. Upon completion of these earlier described components of the interview, the last task participants completed was the RMET. Participant responses were logged using either an Android tablet running ONA, a free and open-source XML-based survey platform, and the Open Data Kit Collect application for interviews that were conducted in the field or using custom software written to automate data collection for interviews conducted virtually and available video conferencing platforms. In either case, conduct of the RMET was comparable and adhered to the procedure as originally described by Baron-Cohen et al (2001), albeit on a digital screen as opposed to paper. Participants were first informed that they would be presented with a series of 36 images of eyes. Each image would feature four words positioned at its corners (translated into and written in the appropriate target language) and the participant would be asked to choose which of the four words they thought best described the emotion or mental state depicted in the image. They were further told that if they were unfamiliar with any of the words, they could freely access a document containing their definitions, synonyms, and usage in a sentence. Participants were told that they could take as long as they

needed but they should try, as much as possible, to move through the items at a quick but comfortable speed. Each item in the RMET constituted a forced choice in which participants were asked to pick the word they felt represented the best fit of the options indicated.

Scoring. Four different coding schemes were used to generate participant scores. The first coding scheme (Unstandardized Baron-Cohen RMET) replicated the original Baron-Cohen et al (2001) coding of which answers were correct. Because methodological artifacts like unfamiliarity with the eye stimuli or the strangeness and rarity of the words could yield mean differences in performance across the languages sampled, a second coding scheme was implemented to correct for this possibility. Thus, in addition to generating raw, Unstandardized Baron-Cohen RMET scores in each field site, Z-scores for each participant, which measured their distance (in standard deviations) from the mean choice of their fellow language speakers were generated. Thus, even if mean performance on the RMET using the Unstandardized Baron-Cohen RMET differed across the three field sites, relationships between RMET performance and amount of MS talk within each language could still be examined. The third coding scheme (Unstandardized Culturally Variable Coding Scheme) was designed to correct for differences across languages in which words are judged most appropriate descriptions for the eye stimuli. This was achieved by scoring as correct those answers that matched the language-specific modal response for each item. As in the case of the Baron-Cohen RMET scores, Z-scores for each participant were also generated here. In this way, distinct words could emerge as the “target” for a given item across the languages sampled while also permitting examination of the relationship between RMET performance and the amount of mental state talk within each of the three languages sampled. Thus, four separate scores were generated for each participant - an Unstandardized Baron-Cohen RMET, a Standardized Baron-Cohen RMET, an Unstandardized Culturally Variable RMET, and a Standardized Culturally Variable RMET

score. Coding the data in these ways allowed the identification of differences across the field sites with respect to response consensus and if so, for which items.

Unstandardized Baron-Cohen RMET Coding Scheme. Participant responses were coded as correct or incorrect according to whether the value selected by the participant matched the target item as originally indicated by Baron-Cohen et al (2001).

Standardized Baron-Cohen RMET Coding Scheme. In addition to their Unstandardized Baron-Cohen RMET scores, Standardized Baron-Cohen RMET were generated within each field site. Unstandardized Baron-Cohen RMET scores within each field site were z-scored to permit observation of where participants from a given field site were positioned relative to other participants from that same field site in the distribution of scores. This approach preserved the ability to observe within-field-site correlations between RMET performance and the production of LR3PMS, even if it were found to be the case that Unstandardized Baron-Cohen RMET scores were incomparable across the three sites sampled.

Culturally Variable Coding Scheme. Within each field site, the highest modal response to each item was calculated and treated as the correct answer for the Culturally Variable coding scheme. Then, a modified version of the criteria applied to each test item by Baron-Cohen et. al. (2001) was employed to identify which of the four words should be treated as the target for that item. Baron-Cohen et. al. decided that items would remain in the test if at least 50% of the participants chose the target word and if no more than 25% of the participants chose any one of the foils. While application of these exact standards was initially intended, in pilot testing it resulted in the removal of several items from the test among the sample of American English speakers. Given this unexpected outcome, the criteria were modified to obviate the dropping of items. Thus, target words for each item in each field site were selected according to whichever had the greatest proportion of participants selecting that word for that item. Using this criterion, three separate coding schemes were generated corresponding to each of the three field sites

sampled. Participants' responses were then evaluated as correct or incorrect according to the coding scheme that corresponded to the field site from which the participant had been recruited. Correct responses were aggregated for each participant, yielding an RMET score that allowed for the possibility of cross-cultural variation in response consensus.

Standardized Culturally Variable Coding Scheme. In addition to their Unstandardized Culturally Variable RMET, Standardized Culturally Variable RMET scores were generated within each field site. Unstandardized Culturally Variable RMET scores within each field site were z-scored to permit observation of where participants from a given site were positioned relative to other participants from that same site in the distribution of scores. This approach preserved the ability to observe within-field-site correlations between RMET performance and the production of LR3PMS, even if it were found to be the case that Unstandardized Culturally Variable RMET Scores were incomparable across the three sites sampled.

Thus, each item to which a participant responded had two "correct/incorrect" values (one of which corresponded to the Unstandardized Baron-Cohen coding scheme and the other of which corresponded to the Unstandardized Culturally Variable coding scheme), and each participant had four RMET scores. The first two scores constituted the total number of items to which their responses matched the values selected by Baron-Cohen et al (2001) and a within-field site standardized score. The second two scores constituted the total number of items to which their responses matched the local consensus and a within-field site standardized score.

LR3PMS and Total Words Uttered Data

Counts of participant LR3PMS were obtained from a corpus of narrative descriptions of video stimuli (See Chapter 2 – General Methods for details). After participant descriptions were transcribed, they were coded for the occurrence of LR3PMS using two different systems. In each case, the count of LR3PMS each participant produced across all of their descriptions was summed to generate the count of LR3PMS to be used in modeling scores on the RMET. This

same summing procedure was performed on the total number of words participants uttered in each of their narrative descriptions, yielding the Total Words Uttered by each participant across all of their narrative descriptions of the video stimuli.

Wellman and Estes Coding. This approach focused on a narrow set of cognitive or belief-like verbs capable of taking sentential complements. These words were “think”, “know”, “believe”, “remember”, “forget”, “mean”, “pretend”, and “dream”. Any instance where one of these verbs referred to a third-party mental state was coded as an LR3PMS.

All Mental States Coding. This approach leveraged the knowledge of coders who were native speakers of each of the target languages. Instead of focusing narrowly on a set of eight mental state verbs, coders identified any and all words that constituted LR3PMS. Coders were provided with a definition of a mental state, and then examined all unique words produced in the corpus of descriptions corresponding to their specific language. Using their knowledge as native speakers and the definition provided to them, coders identified all words that *might* count as LR3PMS. This list was then used to automatically identify them in context, at which point coders did a review of the candidate LR3PMS to remove errors.

Data Analysis

Here, I undertook a procedure of model comparison and selection. Simple linear regression models were analyzed and compared, after which the model with the lowest AIC value was selected for subsequent analysis. This process was undertaken for a total of 8 different datasets, representing each possible combination of the four RMET Scores and the two LR3PMS encodings.

Results

Selecting RMET Measures

The primary goal of this data analysis was to determine whether there existed a correlation between participant performance on the RMET and their production of LR3PMS. As

there were a variety of ways to characterize both performance on the RMET (i.e., Unstandardized Baron-Cohen RMET scores, Standardized Baron-Cohen RMET scores, Unstandardized Culturally Variable RMET scores, and Standardized Culturally Variable RMET scores) and production of LR3PMS (*rate* of LR3PMS, *count* of LR3PMS), checks on data assumptions were first performed (see Appendix C for details). Next, a set of correlation analyses was run to determine whether the choice of particular RMET and LR3PMS measures yielded incongruent, divergent, or substantially differing predictions. To the extent each of the candidate measures correlated strongly and positively with the others, the stronger the evidence of relative invariance in the results regardless of the approach used.

Table 6 contains the correlation coefficients for each of the possible ways to code participant RMET scores and shows that even the most weakly correlated set of predictors was still found to be strongly positively correlated, $r(175) = .9032$, $p < .0001$. Furthermore, no evidence was found to suggest that the mean participant RMET score using the unstandardized Baron-Cohen ($M = 22.232$, $SD = 5.077$) or the Unstandardized Culturally Variable ($M = 23.000$, $SD = 4.835$) coding scheme differed from each other, $t(351.15) = 1.458$, $p = 0.146$. However, it was unclear whether the Unstandardized Baron-Cohen scores were comparable across the three field sites sampled. As there is a growing body of literature to suggest the incomparability of this measure across diverse cultural and linguistic contexts (H. Kim et al., 2022), it was possible that mean scores would differ substantially across field sites. Standardized Baron-Cohen scores would obviate this possibility, though it would limit observation of the relationship between LR3PMS and RMET performance to within field sites. A one-way ANOVA ($F(2, 174) = 2.422$, $p = 0.092$) on the Unstandardized Baron-Cohen RMET scores across field sites provided no evidence of a difference between field sites in mean RMET scores. As such, subsequent analyses employed the Baron-Cohen RMET scores for the sake of consistency with the extant

literature (though see Appendix C for analyses using the Unstandardized Culturally Variable, and Standardized Culturally Variable RMET scores).

Rate of LR3PMS or Distinct Variables for LR3PMS and Total Words Uttered?

All Mental State Terms LR3PMS

Table 7 contains the correlation coefficients for each of the possible ways to characterize the production of LR3PMS and shows that these two measures are only very weakly correlated. Nevertheless, this weak correlation between the *count* of LR3PMS produced by participants and the *rate* at which they produced LR3PMS was found to be statistically significant, $r(175) = 0.1612$, $p = 0.032$, as seen in **Figure 25A**. Though there are good theoretical reasons to believe that the rate at which an individual produces LR3PMS might index their broad orientation toward the minds of others (Carr et al., 2018; Hughes et al., 2018; Meins et al., 2014), the reliability of such an estimate ostensibly depends upon the quantity of speech sampled. Given the relative rarity of content words when compared to function words in speech corpora (Pennebaker et al., 2001, 2003; Tausczik & Pennebaker, 2010), smaller samples of speech may tend to less reliably estimate the rate at which they occur when compared to larger samples. Thus, it was possible that participants who produced fewer total words may thus exhibit greater variance in the estimates of their rate of LR3PMS. Participant *rates* of LR3PMS were plotted against participant *counts* of LR3PMS and against participant word totals, each of which showed that *rates* of LR3PMS exhibited substantially more variation for lower *counts* of LR3PMS and for lower word totals. These data suggest that even if the rate of LR3PMS could be understood as a linguistic index of some underlying orientation to mental states, they cannot be estimated with the same degree of reliability across participants. As such, *counts* of LR3PMS were used in all subsequent analyses. Crucially, it has been claimed by Ruffman et al., (2002) that it is the raw number of LR3PMS, independent of overall speech volume, that matters for the age at which children first pass the False Belief test. However, it is unclear to what extent these variables are

actually dissociable in adult speech. In my sample, these two values were found to be extremely strongly and statistically significantly correlated, $r(175) = 0.929$, $p < .0001$, as seen in **Figure 26A**. For this reason, and for reasons related to the possibility that the *rate* of LR3PMS may still meaningfully predict RMET Scores, word totals were included in all ensuing analyses.

Wellman and Estes Terms LR3PMS

Table 8 contains the correlation coefficients for each of the possible ways to characterize the production of LR3PMS and shows that these two measures are in fact relatively strongly correlated – a feature that differs from All Mental State Terms LR3PMS. Unsurprisingly, this correlation between the *count* of LR3PMS produced by participants and the *rate* at which they produced LR3PMS was found to be statistically significant, $r(175) = 0.4600$, $p < .0001$, as seen in **Figure 25B**. For the same reasons of sampling variance as indicated in the previous section, *counts* and not *rates* of LR3PMS will be used in all subsequent analyses. As in the previous section, a strong and statistically significant correlation was found between the raw number of LR3PMS and the total count of words uttered, $r(175) = 0.4871$, $p < .0001$, as seen in **Figure 26B**. For the reasons indicated above, word totals were included in all subsequent analyses.

Model Comparison and Selection

With analyses of the broad correlations among the data in place, I now turn to construction and comparison of models to find that which best predicts participant performance on the RMET. Per the results described in the previous section, Unstandardized Baron-Cohen RMET scores will serve as the dependent variable and participant counts of LR3PMS will serve as one among potentially many independent variables. Having checked the distributions of the predictors to determine their consistency with the modeling assumptions (see Appendix C for checks of data assumptions), it was clear that Gaussian linear models (as opposed to Poisson or negative binomially-distributed general linear models) would be most appropriate to compare, given the fact that the dependent variable, Unstandardized Baron-Cohen RMET scores, was

itself shown to be effectively normally distributed (see Appendix C). Model fit was determined by AIC. The models compared were chosen to determine if participant performance on the RMET varied across the three field sites sampled, if the count of LR3PMS predicted performance on the RMET, if there existed an interaction between field site and LR3PMS in predicting performance on the RMET, if overall speech volume and count of LR3PMS explained independent elements of the variance in participant RMET performance, and if the count of LR3PMS or overall speech volume more strongly predicted RMET performance. Answers to these questions were thus a matter of identifying the best fit model. **Table 9** provides a detailed view of the models compared. Here I present the results using Unstandardized Baron-Cohen RMET scores. Models using other RMET encodings produced essentially the same results and are thus reported in Appendix C.

Model selection and analysis for Unstandardized Baron-Cohen RMET Scores

LR3PMS Uttered (All Mental State Terms). In all 15 models, simple linear regression was used to determine if the predictor variables significantly predicted Unstandardized Baron-Cohen RMET scores. After running all 15 models, Model 8 was found to have the lowest AIC score (AIC: 1007.141) and predicted Unstandardized Baron-Cohen RMET scores from Total Words Uttered, LR3PMS Uttered, and the interaction between Total Words Uttered and LR3PMS Uttered. Model 8 was found to be statistically significant in its overall fit ($R^2 = 0.3616$, $F(3, 173) = 32.66$, $p < .0001$). It was found that Total Words Uttered ($\beta = 0.007401$, $p < .0001$), LR3PMS Uttered ($\beta = 0.08237$, $p = .0277$), and their interaction ($\beta = -.000067$, $p < .0001$) all significantly predicted Unstandardized Baron-Cohen RMET scores. Thus, for a single-word increase in the count of Total Words uttered, the best-fit model predicted Unstandardized Baron-Cohen RMET performance to increase by approximately 0.007 points. As the RMET is scored on an integer scale ranging from 0 to 36, it is perhaps more meaningful to frame this result in terms of the change in Total Words Uttered expected to correspond to a single-point increase in

Unstandardized Baron-Cohen RMET score. Characterized thusly, a single-point increase in Unstandardized Baron-Cohen RMET score is expected for each increase of 136 words in Total Words Uttered. For a single-word increase in the count of LR3PMS uttered, the best fit model predicted Unstandardized Baron-Cohen RMET scores to increase by approximately 0.08 points.

As before, this means that for a 13 word increase in the count of LR3PMS uttered, Unstandardized Baron-Cohen RMET score was expected to increase by approximately 1 point. Finally, the interaction term can be understood as follows. For a single-word increase in the count of Total Words Uttered, each single-word increase in the count of LR3PMS uttered was expected to *decrease* Unstandardized Baron-Cohen RMET scores by 0.0007 points. **Figure 27A** illustrates the predicted Unstandardized Baron-Cohen RMET scores with 95% confidence intervals across the range of values for Total Words Uttered. **Figure 28A** illustrates the predicted Unstandardized Baron-Cohen RMET scores with 95% confidence intervals across the range of values for LR3PMS. **Figure 29A** illustrates the effect of LR3PMS Uttered on Unstandardized Baron-Cohen RMET scores with 95% intervals for participants with a Total Words Uttered count of 395 (the value corresponding to the lower quartile of words uttered), 790 (the median number of words uttered), and 1261 words (the value corresponding to the upper quartile of words uttered). **Figure 30A** is essentially the same as **Figure 29A**, though the x-axis now corresponds to the count of Total Words Uttered. **Figure 30A** illustrates the effect of Total Words Uttered on Unstandardized Baron-Cohen RMET scores with 95% intervals for participants with an LR3PMS count of 15 (the value corresponding to the lower quartile of LR3PMS uttered), 31 (the median number of LR3PMS uttered), and 55 words (the value corresponding to the upper quartile of LR3PMS uttered). Figures corresponding to the fit statistics of Model 8 can be found in Appendix C.

LR3PMS Uttered (Wellman and Estes Terms). In all 15 models, simple linear regression was used to determine if the predictor variables significantly predicted

Unstandardized Baron-Cohen RMET scores. After running all 15 models, Model 8 was found to have the lowest AIC score (AIC: 1010.005) and predicted Unstandardized Baron-Cohen RMET scores from Total Words Uttered, LR3PMS Uttered, and the interaction between Total Words Uttered and LR3PMS Uttered. Model 8 was found to be statistically significant in its overall fit ($R^2 = 0.3512$, $F(3, 173) = 31.21$, $p < .0001$). It was found that Total Words Uttered ($\beta = 0.007451$, $p < .0001$), LR3PMS Uttered ($\beta = 0.653796$, $p = .0109$), and their interaction ($\beta = -.000073$, $p < .001$) all significantly predicted Unstandardized Baron-Cohen RMET scores. Thus, for a single-word increase in the count of Total Words uttered, the best-fit model predicts Unstandardized Baron-Cohen RMET performance to increase by approximately 0.007 points. As the RMET is scored on an integer scale ranging from 0 to 36, it is perhaps more meaningful to frame this result in terms of the change in Total Words Uttered expected to correspond to a single-point increase in Unstandardized Baron-Cohen RMET score. Characterized thusly, a single-point increase in Unstandardized Baron-Cohen RMET score was expected for each increase of 136 words in Total Words Uttered. For a single-word increase in the count of LR3PMS uttered, the model predicted Unstandardized Baron-Cohen RMET scores to increase by approximately 0.70 points.

As before, this means that for a 2 word increase in the count of LR3PMS uttered, Unstandardized Baron-Cohen RMET score was expected to increase by approximately 1 point. Finally, the interaction term can be understood as follows. For a single-word increase in the count of Total Words Uttered, each single-word increase in the count of LR3PMS uttered was expected to *decrease* Unstandardized Baron-Cohen RMET scores by 0.0007 points. **Figure 27B** illustrates the predicted Unstandardized Baron-Cohen RMET scores with 95% confidence intervals across the range of values for Total Words Uttered. **Figure 28B** illustrates the predicted Unstandardized Baron-Cohen RMET scores with 95% confidence intervals across the range of values for LR3PMS. **Figure 29B** illustrates the effect of LR3PMS Uttered on Unstandardized

Baron-Cohen RMET scores with 95% intervals for participants with a Total Words Uttered count of 395 (the value corresponding to the lower quartile of words uttered), 790 (the median number of words uttered), and 1261 words (the value corresponding to the upper quartile of words uttered). **Figure 30B** is essentially the same as **Figure 29B**, though the x-axis now corresponds to the count of Total Words Uttered. **Figure 30B** illustrates the effect of Total Words Uttered on Unstandardized Baron-Cohen RMET scores with 95% intervals for participants with an LR3PMS count of 1 (the value corresponding to the lower quartile of LR3PMS uttered), 3 (the median number of LR3PMS uttered), and 4 words (the value corresponding to the upper quartile of LR3PMS uttered). Figures corresponding to the fit statistics of Model 8 can be found in Appendix C.

Discussion

In the current study, I compared a set of 15 distinct models featuring various combinations of the following predictors: Total Words Uttered; LR3PMS Uttered; Field Site. This process was done twice – once for an encoding of LR3PMS that included a set of only mental state verbs capable of taking sentential complements and once for an encoding of LR3PMS that included all mental state terms. Each of the models tested corresponded to a distinct set of answers to the questions I aimed to address in the current study. Models in the set featured at least one of these three predictors, though most were more complex than simple univariate regression models. Here, I report the results of the model comparison wherein Unstandardized Baron-Cohen RMET scores were predicted from All Mental State Terms LR3PMS, as well as the results of the model comparison wherein Unstandardized Baron-Cohen RMET scores were predicted from Wellman and Estes LR3PMS. Importantly, the best fit model in both of these model comparison procedures corresponded to Model 8 as specified in **Table 9**. Results for all other RMET and LR3PMS encodings can be found in Appendix C.

First, I found that participant performance on the RMET did not vary across field sites. Regardless of the LR3PMS and RMET encodings I used, the best fit model (Model 8) predicted RMET scores from just the count of Total Words Uttered by participants, the count of LR3PMS Uttered by participants, and the interaction of these two terms. Field Site did not explain a meaningful degree of variance in RMET scores between American, Chinese, and Moroccan Participants. I also found that regardless of which of the four encodings of RMET scores I used (Unstandardized Baron-Cohen, Standardized Baron-Cohen, Unstandardized Culturally Variable, Standardized Culturally Variable), no statistically significant differences were observed across groups in mean participant score.

Second, I found that those participants who spoke more in the course of their elicited video descriptions, regardless of how many LR3PMS they produced, tended to attain higher scores on the RMET. This finding was effectively identical across both LR3PMS encodings. More specifically, it was found that Total Words Uttered strongly positively predicted Unstandardized Baron-Cohen RMET score. The effect of a single word increase in the count of Total Words Uttered on Unstandardized Baron-Cohen RMET score was actually smaller than that of a single word increase in LR3PMS; however, the range of values represented in Total Words Uttered was substantially greater than that of LR3PMS Uttered. As such, the expected increase in Unstandardized Baron-Cohen RMET score for a single-word increase in Total Words Uttered could be smaller than that of a single-word increase in LR3PMS Uttered while nevertheless constituting an otherwise stronger predictor of RMET scores. Thus, **Figures 27A and 27B** illustrate the strength of this relationship when holding the count of LR3PMS Uttered constant at the sample mean.

Third, I found that those participants who produced a greater number of LR3PMS, regardless of how many words they uttered overall, tended to attain higher scores on the RMET. However, the strength of this effect at the mean value of Total Words Uttered appeared to differ

depending on the specific encoding of LR3PMS such that the relationship was stronger (i.e., exhibited a larger effect size and was statistically significant at a lower alpha value) when coding for only the Wellman and Estes terms. That is to say, LR3PMS Uttered positively predicted Unstandardized Baron-Cohen RMET scores, albeit more weakly than Total Words Uttered. As can be seen in **Figure 28A**, there was a weak increase in RMET Score associated with a greater number of All Mental State Terms LR3PMS Uttered when holding Total Words Uttered constant at the sample mean. Curiously, **Figure 28B** illustrates that there was a slight *decrease* in RMET score associated with a greater number of Wellman and Estes Terms LR3PMS Uttered when holding Total Words Uttered constant at the sample mean. These visualizations are a bit puzzling given the fact that the independent effect of each of these predictors was quite strongly positive. However, this finding can be understood in light of the fact that the interaction between Total Words Uttered and LR3PMS *negatively* predicted Unstandardized Baron-Cohen RMET scores, which relates to my next finding.

Fourth, I found that the more talkative a participant was, the less the count of LR3PMS they uttered mattered in predicting their score on the RMET (see **Figures 29A and 29B** as well as **Figures 30A and 30B**). Among the most talkative participants, scores on the RMET actually decreased as the number of LR3PMS uttered increased. In contrast, the least talkative participants saw increases in their RMET scores as their counts of LR3PMS uttered increased.

Finally, I found that the amount of variance in RMET scores explained by participant counts of LR3PMS was effectively identical across the All Mental States LR3PMS encoding and the Wellman and Estes LR3PMS encoding. When using All Mental State Terms LR3PMS as a predictor in the model comparison procedure, the best fit model explained 36.16% of the variance in Unstandardized Baron-Cohen RMET scores. When using Wellman and Estes Terms LR3PMS as a predictor, the best fit model explained 35.12% of the variance in Unstandardized Baron-Cohen RMET scores. In both cases, all three predictors in Model 8 were statistically

significant, of the same relative magnitude, and in the same direction. Thus, the All Mental States LR3PMS encoding accounted for approximately a single percentage point more of the variance in participant RMET scores despite increasing mean participant LR3PMS counts from approximately 3 to 36. A tenfold increase in the mean value of this independent variable accounted for effectively no new variation in RMET scores, suggesting that these two schemes are redundant in the information they encode. This increase did correspond to a comparable reduction in the effect size of the LR3PMS count, such that the beta value when LR3PMS count is encoded using the All Mental States scheme is approximately a tenth of the value when LR3PMS count is encoded using the Wellman & Estes scheme.

These findings are notable given that no model featuring Field Site as a predictor emerged as best fit across all eight combinations of LR3PMS and RMET score encodings, even though Field Site featured as a predictor in 11 of the 15 models tested. Though the AIC values of the second, third, fourth, and fifth best fit models were comparable and these models did include Field Site as a predictor, Chi-square goodness of fit tests comparing Model 8 to each of these runners-up failed to show a statistically significant improvement of fit with the additional variables they contained. The fact that neither Field Site alone, nor its interactions with the other predictors, emerged in the best fit model, regardless of how LR3PMS and RMET scores were encoded, suggests two separate conclusions. The first is that participant RMET scores themselves did not vary meaningfully across the three field sites from which the data was collected. The second is that the relationships between participant RMET Scores, the Total Words Uttered by participants, the LR3PMS Uttered by participants, and the interaction of these two variables did not vary across the three field sites from which the data was collected either. Notably, these runners-up were the same models regardless of RMET Score and LR3PMS encoding (Model 11, Model 12, Model 14, and Model 15), though their specific ordering varied. As before, this finding held across all eight combinations of LR3PMS and RMET Score

encodings. Even though the RMET has been shown previously to generalize poorly across cultural and linguistic contexts (H. Kim et al., 2022), these results suggest that statistically significant differences in scores across Field Sites do not necessarily emerge when using the RMET as originally designed and scored by Baron-Cohen et al., (2001).

These findings complicate the hypothesis that there exists a straightforward relationship between RMET performance and the quantity of mental state talk, though they also provide preliminary evidence to suggest this phenomenon manifests similarly across the diverse cultural and linguistic contexts sampled. Moreover, these findings raise questions about the role of participant talkativeness, as those who spoke more in the aggregate attained higher RMET scores. Crucially, the meaning and interpretation of these results will depend upon adjudicating whether they should be understood as constituting real effects or methodological artifacts. In the following section, I first contextualize these findings in light of the questions this study aimed to answer. I then detail possible interpretations of these findings conditional on the assumption that they constitute real effects not attributable to methodological limitations. Then, I detail caveats on these interpretations attributable to the limitations of the methodologies employed herein.

Interpreting the Findings and Implications for Extant Literature

Firstly, these findings support the conclusion that, at least for the three field sites I sampled, the RMET may be sufficiently generalizable in its application across cultural and linguistic contexts to permit comparison against those sampled here. The finding that there is no effect of Field Site on RMET Score, nor an interaction effect of Field Site with any of the other predictors in the model suggests that these findings may be representative of a phenomenon that instantiates more or less similarly across human populations. In effect, though many features of American English, Mandarin Chinese, and Moroccan Arabic differ, the relationship between the production of mental state talk and mindreading may be language invariant. Though criticisms have been levied against the RMET with respect to its cross-cultural and

cross-linguistic generalizability, it is perhaps possible that cultural and linguistic contexts vary with respect to the applicability of the measure as it was originally designed. For example, societies with greater exposure to foreign media or highly ethnically and racially diverse population centers may be sufficiently equipped to make emotion and mental state attributions to individuals belonging to social categories distinct from one's own. This challenge is one that has been documented previously (Adams Jr et al., 2010; Bjornsdottir & Rule, 2016; H. Kim et al., 2022). Beyond what these findings indicate about the efficacy of the RMET in diverse linguistic and cultural contexts, the fact that no effect of Field Site was observed is consistent with the notion that there are universal emotional categories and broad homogeneity in the capacity to attribute mental states (Ekman et al., 1987; Jackson et al., 2019; Wellman, 2013). Though the Culturally Variable coding scheme allowed the "target" word to differ across field sites for all 36 items, the target words of most items remained unchanged from the original Baron-Cohen coding scheme. However, these data also suggest that the specific meaning attributed to particular facial expressions and other related social perceptual stimuli may in some cases vary across cultures, as has been documented previously by emotion researchers (Aival-Naveh et al., 2019; Lillard, 1998). That is, the "target" word for a number of the RMET items differed across cultures when using the Culturally Variable coding schemes. Nevertheless, these variable targets were generated according to whichever of the four words was most frequently selected by participants in a given field site. Thus, these data suggest that cultural and linguistic frameworks may condition the meaning of at least some social stimuli. Though the meaning may change, individuals within a given cultural or linguistic context vary in the extent to which their attributions of others' mental states align with those of the culturally-determined consensus. These findings are consistent with corpus linguistic research which has suggested there is both uniformity and diversity in emotion terms across languages (Gendron et al., 2014; Jackson et al., 2019; Matsumoto, 1989). These findings suggest that the relationship between

the production of mental state talk and underlying mindreading competence may be less sensitive to cross-cultural or cross-linguistic differences than has been implied by findings in psychology and anthropology (Bradford et al., 2018; J. G. Miller, 1986).

Next, I contextualize these findings in terms of the hypothesis and predictions enumerated in the introduction of this chapter. First, performance on the RMET is just as effectively predicted by the count of belief-like mental state verbs as it is by the production of any and all mental state terms. That is to say, no additional information appears to be captured by the inclusion of words beyond just those for belief-like mental state verbs. This finding is consistent with claims that have been made in the developmental literature proposing a special role of cognitive or belief-like verbs in supporting the development of mindreading (Cheung et al., 2004; de Villiers, 2005; de Villiers & Pyers, 2002; Gleitman, 1990). Crucially, no analyses of the syntactic properties of the Wellman and Estes Terms LR3PMS across Mandarin, American English, and Moroccan Arabic were performed. As such, these findings can neither confirm nor disconfirm the hypothesized role of the syntactic properties of these verbs. It should also be noted that the theories which posited a role of cognitive or belief-like verbs were primarily concerned with their influence on mindreading in early childhood and were agnostic as to whether longer-term impacts of differential exposure to and mastery of such verbs existed.

That there remains a predictive relationship between participants' production of such terms and their mindreading ability potentially extends the ontogenetic scope of these theories into adulthood. Though these findings are preliminary, recent research has extended the range of ages for which there are purported relationships between the production of mental state talk and mindreading ability. In a recent longitudinal study, Carr et al. (2018) found that children's production of mental state language (glossed in their study as including cognitive terms, desire terms, emotion terms, general mental state terms, and modulations of assertion) and their performance on a battery of mindreading tasks was moderately correlated (r=0.40) for three

year old children, though this correlation disappeared for the same children at ten years of age. It should be noted, however, that the mindreading tasks implemented at three years of age and ten years of age differed substantially in the extent to which participant responses had the potential for floor effects. Whereas 95% of participant scores for the tests administered at age three covered effectively the full range of potential values (0 – 5), scores on the test administered at age ten could range from 0 to 24. 95% of participant scores ranged between 10 and 22 points. This reduction in variance may limit the ability to detect a relationship that might otherwise hold, as suggested by the present findings. It warrants mention that participant production of mental state terms was moderately correlated across the three-year and ten-year data, indicating that this quality may be a stable linguistic behavior across development. These authors also reported stability in maternal mental state talk across the same period of 7 years in their sample of adults producing child-directed speech.

Another assumption built into the current study's predictions was that the relationship between the production of belief-like mental state verbs and False Belief Test performance would generalize to other measures of mindreading – in this case, the RMET. As effectively no difference in the percentage of variance explained by the best fit model across each encoding of LR3PMS Uttered was found, it seems that the frequency of production of cognitive or belief-like verbs is at least as meaningful a predictor of emotion detection as is the frequency of all mental state terms. Though it had been reported previously that the production of such terms predicted performance on the False Belief Test, it was not necessarily the case that they would predict performance on the Reading the Mind in the Eyes Test, a distinctly less epistemically focused measure of mindreading abilities. If the semantic properties of the words mattered more than the syntactic ones for mastery of a given mindreading subdomain, it seemed plausible that the All Mental States Terms LR3PMS would account for a greater degree of variance. However, no such difference was found. One possibility is that because emotions and desires are mental

state representations that emerge earlier in development, they may constitute less complex representational types. For example, they may be thought of as first-order representations, that do not require one to represent another's representation. I might represent *that* you are happy, but I might not represent *why* you are happy. The usage of such terms may thus tax the mindreading system in a less obligately strong way than the use of more cognitive terms that represent second-order mental states (Leslie & Happé, 1989; S. A. Miller, 2009; Sullivan et al., 1994). This aligns with claims in the literature about desire-like representations emerging earlier in development than belief-like representations, such as children mastering the concept of others wanting things before mastering that they think differently from themselves (Avis & Harris, 1991; Harrigan et al., 2018; Wellman & Liu, 2004).

I next turn toward the finding that overall speech volume predicted performance on the RMET as strongly, if not more so, than the count of LR3PMS. It was predicted that participant production of LR3PMS would be the only variable to account for participant RMET performance. This prediction was predicated on a finding reported by Ruffman et al., (2002) showing that the age at which children first passed the False Belief Test was predicted by the raw count of mental state terms uttered by their parents, and not the relative frequency of such terms. In effect, both a taciturn mother and a gregarious mother who uttered 10 mental state terms could expect their children to pass the False Belief Test at the same age. Consequently, I predicted that the best fit model would not feature Total Words Uttered. This, however, was not the case. Both a main effect of participant verbosity, as measured by Total Words Uttered, and its interaction with LR3PMS Uttered were significant predictors of participant RMET scores in the best fit models. Taking these results at face value, one possible interpretation is that across diverse cultural and linguistic contexts, people who are more efficient or more accurate mind readers may find social interactions easier, a possible consequence of which is increased speech duration or fluency when interacting with novel social partners (as might be the case when providing narrative

descriptions of video stimuli to an unfamiliar experimenter). Under such an interpretation, a greater volume of speech indirectly indexes the absence of something like social anxiety. In contrast, a participant who is less effective at understanding the mental states of their interlocutors may have greater social anxiety owing in part to some uncertainty about their reception by their interlocutors. As such, speech is reduced. This account has some empirical backing – in a recent meta-analysis, Baez et al. (2023) found that compared to neurotypical controls, individuals with social anxiety disorder exhibited impairments in both emotion recognition and mental state attribution. Both emotion recognition and mental state attribution were measured using a number of tests, including the RMET Test (Baron-Cohen et al., 2001), the Movie Assessment of Social Cognition, or MASC (Dziobek et al., 2006), and the Faux Pas Test (Baron-Cohen et al., 1999). Furthermore, Scharfstein et al. (2011) demonstrated that socially-phobic children have poorer overall social skills than neurotypical or autistic children, including greater latency to speak, fewer words uttered, inappropriate affect, inappropriate responses in conversational turn-taking, and lower effort to maintain conversations.

This account emphasizing the lack or presence of social anxiety is, notably, just one among many possible explanations of the data. For example, greater speech production may in fact represent an overall facility with spoken language production that may translate to other traits like vocabulary size and performance on tests involving written or spoken language. Another plausible explanation of these findings is that individuals who find social interaction more intrinsically rewarding may interact with others more frequently and with greater duration, exposing them to interlocutors' mental states more frequently and improving the accuracy with which they impute them. This hypothesis, known as the Social Motivation Theory (Chevallier et al., 2012), is one for which there is some evidence. Bagg et al. (2024) recently demonstrated that in a sample of 165 adolescents, approximately half of whom were neurotypical and half of whom were autistic, individuals with higher social motivation exhibited fewer and less intense

autistic traits. Social motivation was measured using the Choose-A-Movie paradigm which had been used previously to measure participants' effort to view social and nonsocial stimuli, with autistic participants showing a greater preference for nonsocial stimuli (Dubey et al., 2015).

Finally, I turn to the interaction between Total Words Uttered and LR3PMS Uttered. These results suggest the extent to which the count of LR3PMS Uttered predicted performance on the RMET depended upon the overall volume of speech produced. Those who spoke relatively little tended to score better as the count of LR3PMS they uttered increased. Those who spoke a great deal, however tended to score more poorly as the count of LR3PMS they uttered increased. This finding is puzzling given the claims made by Ruffman et al., (2002), especially in light of the strength of the correlation between LR3PMS and Total Words Uttered. If the value of LR3PMS in predicting RMET scores varies as a function of overall speech volume, there may be distinct behavioral or psychological processes driving overall talkativeness with divergent effects on the count of LR3PMS uttered. Thus, focusing on these variables independently collapses distinct behavioral or psychological profiles.

For example, individuals with high degrees of social anxiety may withdraw socially and thus produce fewer words overall. However, they may also struggle with appropriately attributing mental states to others. Such individuals might produce relatively few words overall, and of those produced, very few may be LR3PMS. They might similarly perform more poorly on the RMET. In contrast, individuals who are not socially anxious but are less gregarious may not struggle at all with attributing mental states to others. Given the pragmatic demands of the video description task, the proportion of words they utter that constitute LR3PMS might consequently be higher (assuming that the inclusion of LR3PMS is what the pragmatic context demands) and they might perform more strongly on the RMET. A similar kind of logic may apply on the opposite end for highly gregarious individuals. This might be thought of as a kind of social error management strategy (Haselton & Buss, 2000) that covaries with gregariousness. Those who

are highly gregarious but poor mind readers may produce relatively more LR3PMS, a greater proportion of which are “false positives”. Though these individuals speak a great deal, it is possible that the greater frequency of LR3PMS is in fact a kind of linguistically mediated opportunity to fact check their attributions of mental states. This behavior would then belie the same skill that accounts for poorer performance on the RMET. In contrast, those who are more taciturn and also poor mind readers may decrease the relative frequency of LR3PMS, missing genuine or important mental states borne by their interlocutors and thus having a greater proportion of “false negatives”. In either case, such individuals may be less accurate or less efficient mind readers. Taciturn yet effective mind readers may nevertheless commit more “false negatives” errors than their more gregarious counterparts by virtue of having less to say overall. However, the more effective a mind reader they are, the greater the number of LR3PMS they may produce and the higher they may score on the RMET. Gregarious and effective mind readers may also commit more “false positives” of mental state attribution than their more taciturn counterparts, but beyond establishing the cognitive “facts” of the matter, their greater volume of speech may instead track other aspects of interactional style. These skills may predict both speech volume and performance on the RMET. Future research may be well-served to explore the possibility of this hypothesis, as this negative interaction is a puzzling result that throws a wrench in what might otherwise be a straightforward set of relations between the variables tested here.

Caveats and Limits on Interpretation

It is not strictly a given that the results reported here should be taken at face value. The interpretations reported above are conditional on the validity of the measures employed. However, there are at least some reasons to treat them with caution. As indicated above, the RMET has been subject to intense criticism with respect to its generalizability across cultural and linguistic contexts, as well as with respect to its psychometric validity (Black, 2019). Item

response theorists have shown that the test is riddled with threats to its validity. Though measures were undertaken in the present study to address the first criticism, Z-scoring participant responses cannot necessarily account for flaws of the type indicated by the second criticism. Additionally, the ways in which emotion perception actually differs cross-culturally is not well-established, if indeed it differs at all (Ekman, 1992; Elfenbein & Ambady, 2002; Sauter et al., 2010). There is evidence for both universality and cross-cultural variation in the set of emotions and facial expressions experienced and produced by human beings (Ekman et al., 1987; Gendron et al., 2014; Sauter et al., 2010). Even if they are universal, it is not guaranteed their linguistic glossing is comparable cross-linguistically.

Furthermore, there are potential drawbacks of the methods employed to collect participant speech. Though the video stimuli used were designed to be minimally culturally-laden and applicable across a wide range of cultural and linguistic contexts, they were not psychometrically validated. As such, it is possible that the variation observed did not capture individual differences in the propensity to produce LR3PMS, or that it failed to do so in a way that was equivalent across field sites. Moreover, it is an assumption that the elicited speech samples collected were representative of participants' speech outside of the research context. Though there is some evidence to suggest a contribution of personality to social behavior across various contexts (Berge & Raad, 2001; Murtha et al., 1996; Pennebaker & King, 1999), it was unknown whether this contribution was of the same magnitude for all participants. Thus, these samples may not be strictly representative of participants' general tendency to talk about the minds of others. Additionally, though there was effectively no difference in the variance of RMET scores explained by the best fit model across the two encodings of LR3PMS Uttered, an important caveat is that the Wellman and Estes Terms encoding did not cleanly capture *every* cognitive or belief like verb. Instead, it captured only those 8 that were reported to be the most common in children's speech per Wellman and Estes (1987). Therefore, every other cognitive or

belief like verb produced in the corpus would not be included in the Wellman and Estes coding scheme, though it would be including in the All Mental State Terms encoding. To the extent that the All Mental State Terms encoding captures predominantly synonyms of the Wellman and Estes terms or other cognitive and belief-like verbs, these different approaches may in fact be measuring the same phenomenon and as such, do not allow accurate observation of the effect of other categories of mental state terms. Additionally, the finding that Total Words Uttered predicts performance on the RMET may not be directly attributable to the quantity of speech, but to some third, associated variable like vocabulary size or general linguistic facility. If so, participants who produce more speech in the elicited narrative description task may just have a greater productive and receptive vocabulary. As such, the choices participants make in the RMET can be more confidently attributed to the accuracy of their identification of the item, as they may have greater familiarity with the target word and the three foils. Participants with smaller vocabularies may speak less and be less familiar with the words associated with each RMET item. As such, it can be less confidently stated that their performance is actually associated with their ability to recognize mental states in others. Instead, it may be that they select the best word amongst those with which they are familiar. Future investigations will be vital to discount this possibility.

Conclusion

In summary, the approach implemented in this study allowed many, though not all, of the outstanding questions identified at its outset to be addressed. The results show a relationship between the production of LR3PMS and mindreading ability, as measured by the RMET, among adults. An independent relationship of total speech volume with performance on the RMET was also demonstrated. Notably, a negative interaction between these two factors was observed with respect to predicted RMET scores. These findings were then contextualized in the extant literature on human universal and cross-cultural variation. Additionally, preliminary evidence of a

unique relationship between mindreading ability and belief-like verbs, as opposed to mental state terms more generally was reported. The findings also provided evidence to suggest there is not a domain-specific effect of type of mental state word on different subdivisions of the mindreading capacity. Finally, these findings illustrated it was the raw count, and not the relative frequency of LR3PMS that best predicted performance on the RMET. These results have important implications for contemporary understanding of the relationship between mindreading and the production of mental state talk, as well as between mindreading and linguistic behavior construed generally.

Chapter 6: General Conclusion

Introduction

This dissertation set out to investigate whether mental state talk varies across three distinct cultural-linguistic contexts and to provide initial insight into how these patterns of mental state talk are related to underlying mindreading ability among neurotypical adults. Through the examination of lexical references to third-party mental states (LR3PMS) and their correlation with mindreading performance, I aimed to address longstanding theoretical debates about the universality and variability of mental state talk, as well as to challenge and complexify some of the functional claims that have linked exposure to mental state talk to the development of theory of mind. By systematically analyzing these relationships across multiple studies, the findings presented here provide new insights into cross-cultural dimensions of social cognition and the role of language in shaping our understanding of other minds. Thus, this work elucidates further the complex relationship between mental state talk and mindreading ability across various cultural-linguistic contexts.

Chapter 1 laid important theoretical groundwork by reviewing existing literature and identifying key gaps in our current understanding of how mental state talk varies across languages and cultures, how mindreading varies across languages and cultures, and how these constructs have been shown to relate to each other. Chapter 2 introduced a novel methodology by which to generate standardized corpora of speech samples designed with the express goal of comparing the production of mental state talk across languages. This methodology was then employed to generate a cross-linguistic corpus of speech samples generated by English-speaking participants recruited from the United States, Mandarin-speaking participants recruited from China, and Arabic-speaking participants recruited from Morocco. These data were subsequently employed in empirical investigations presented in Chapters 3, 4, and 5. These empirical studies each addressed specific aspects of the relationship between mental state talk

and mindreading. Chapter 3 examined whether the production of a narrow set of cognitive or belief-like verbs (Wellman and Estes Terms), the importance of which to mindreading development had been posited previously in the literature (de Villiers, 2005; de Villiers & Pyers, 2002; Gleitman, 1990), differed across the three field sites sampled. Chapter 4 identified potential flaws in such a limited conception of mental state terms and therefore leveraged the emic knowledge of native speakers to code the data for any and all LR3PMS (All Mental State Terms). Chapter 5 then took these data to examine whether participants' production of LR3PMS across these two coding systems predicted their scores on the Reading the Mind in the Eyes Test (RMET), a widely used measure of mindreading ability in adults, as well as to see whether this relationship differed across the three field sites sampled.

Summary of Key Findings

The central finding of this dissertation is that whether mental state talk varies cross-linguistically depends upon how mental states are defined. When focusing on just propositional attitudes (as represented by the Wellman and Estes coding scheme employed in Chapter 3), which are lexicalized in English through cognitive or belief-like verbs capable of taking sentential complements, no cross-linguistic differences in the absolute frequency of their production in speech was observed. However, substantial differences were observed in the relative frequency of such mental states. In contrast, the opposite pattern was found when focusing on a broader range of internal states available to an individual's conscious awareness (as represented by the All Mental States coding scheme employed in Chapter 4). Significant differences were observed in the absolute frequency of their production in speech, but their relative frequency was essentially the same across the three cultural-linguistic groups sampled. Because description lengths of the videos varied across the three cultural-linguistic groups, these findings suggest that the frequency of propositional attitudes may be dissociable from overall speech quantity whereas the frequency of talk about any and all internal states may instead tend to represent a

relatively fixed percentage of overall speech, regardless of the language spoken by an individual (Goddard, 2010; Jackson et al., 2019). In effect, while the absolute counts of LR3PMS did not differ significantly across the three field sites when using a conservative coding scheme (Chapter 3), broadening the scope of what constitutes mental state talk revealed cross-cultural differences in absolute counts (Chapter 4). This suggests that cultural and linguistic factors influence the specific ways in which people talk about the minds of others, which has important implications for theories that posit a universal link between mental state talk and mindreading development.

Another important finding presented in Chapters 3 and 4 was that the primary determinant of how frequently participants produced LR3PMS was the video stimuli themselves, with the field site from which participants were recruited consistently accounting for less variance in the count of LR3PMS. The situations depicted in the videos varied substantially in how strongly they elicited mental state talk, with some stimuli consistently generating high or low counts of LR3PMS across all three cultural contexts and others showing more cross-linguistic variability. This result underscores the significant role played by contextual factors in eliciting mental state talk, suggesting that even if there do exist cross-cultural differences in mental state talk, it is less likely to be a product of inherent cultural-linguistic differences than it is to be a product of the kinds of social interactions encountered and spoken about by individuals across cultural-linguistic contexts. As such, this finding raises important questions about the differences in the patterning of daily social life across cultural-linguistic contexts and whether they entail differential amounts of mental state talk. It is transparently the case that contextual factors influence the content of individuals' speech (W. S. Hall et al., 1981; Parrigon et al., 2017; Sherman et al., 2015). If, however, there are systematic differences across societies in the duration or frequency with which a given context is encountered, then there may too be differences in the frequency of mental state talk. While previous work has examined language in

“everyday life” (Campos et al., 2009; Ochs et al., 2011), it is important to know if these patterns hold with respect to the universality of specific contextual influences on linguistic behavior and the universality of the time spent within those specific contexts.

Chapter 5 presented several new insights. The first was that mean scores on the RMET across field sites did not differ, regardless of whether they were scored using Baron-Cohen or Culturally Variable coding schemes. The second was that the best fit model to explain participants’ scores on the RMET included the count of LR3PMS Uttered by participants, the count of Total Words Uttered by participants, and the interaction between these two factors as predictors. This same model constituted the best fit regardless of whether LR3PMS were coded using the Wellman and Estes Terms coding scheme employed in Chapter 3 or the All Mental States Terms coding scheme employed in Chapter 4. The results presented in Chapter 5 complicate our present understanding of the relationship between mindreading and language, as participant Scores on the RMET were independently predicted by both the count of LR3PMS and the total number of words uttered by participants. This finding challenges the notion that it is only the presence of specific mental state verbs that matters for mindreading competence. Instead, it suggests that overall speech quantity plays a dissociable role in mindreading performance. This suggestion is consistent with the negative interaction that emerged in my best fit models. Effectively, the greater the total number of words a participant uttered, the more negatively the total number of LR3PMS they uttered predicted their RMET. Thus, the most loquacious participants actually saw reductions in their RMET scores as the total number of LR3PMS they uttered increased. This finding complicates claims made by Ruffman et al., (Ruffman et al., 2002) and suggests that dissociating the frequency of mental state talk from overall speech quantity may obfuscate important variation in mindreading phenotypes (Baez et al., 2023; McCroskey & Richmond, 1995; Scharfstein et al., 2011). A final and notable finding was that the variation in participant RMET scores explained by my statistical model was

effectively the same across the two encodings of LR3PMS. This finding suggests that counting all references to internal states, as opposed to just cognitive or belief-like verbs, did not serve to improve the explanatory power of the model.

Theoretical Implications

One of the primary contributions of this dissertation is its challenge to the oversimplified view that mental state talk does not vary across cultural-linguistic contexts and that it is a direct and universal predictor of mindreading ability. This has significant implications for theories of theory of mind development that emphasize the importance of specific linguistic structures, such as cognitive or belief-like verbs that can take sentential complements. The Wellman and Estes Terms coding scheme used in Chapter 3, which focused on such verbs, revealed no significant cross-cultural differences in the absolute counts of LR3PMS, implying a level of universality in the use of these structures. However, Chapter 4's All Mental State Terms coding scheme, which included a wider range of mental state terms, found cross-cultural differences in the absolute counts of LR3PMS. Given the ambivalence of the findings presented across Chapters 3 and 4, the findings presented in Chapter 5 may be understood as shedding important light on claims made in the literature pertaining to the relationship between mental state talk and mindreading ability. Given no meaningful improvement was observed in the explanatory power of the best fit model predicting RMET from LR3PMS across its encodings, the additional words captured by the All Mental States scheme may be less strongly related to mindreading competence than those captured by the Wellman and Estes scheme.

If so, this has implications for claims that have been made about cross-cultural variation in mental state talk. If there are no absolute differences in the production of cognitive or belief-like verbs across populations, and these are, in fact, the kinds of mental state terms that predict underlying mindreading competence, then the data presented here suggest it is unlikely that underlying mindreading competence would vary cross-linguistically as a function of the

frequency of mental state talk. Though production of any and all mental state terms may vary across languages, those that have been purported to play a role in mindreading development (Cheung et al., 2004; de Villiers, 2005; de Villiers & Pyers, 2002; Gleitman, 1990) and that may serve as an index of mindreading ability occur at effectively the same frequency across cultural-linguistic contexts. One caveat on this interpretation is the fact that there exist differences in mean number of words uttered per transcript across the three cultural-linguistic groups sampled. If these patterns are indicative of broader differences in overall talkativeness across these groups, then the interpretation provided in the previous paragraph may hold. However, if these differences in talkativeness are simply an artifact of sampling, then this interpretation should be treated with caution, as the relative count of cognitive or belief-like verbs varied significantly across the three sites. Collectively, these data suggest that focusing solely on cognitive or belief-like verbs may obscure variation in the production of all mental state terms, but whether such variation has functional consequences remains unclear. The findings reported here suggest that those terms which have been posited to have a functional role in the development of the mindreading capacity occur at the same frequency and predict RMET performance in similar ways across cultural-linguistic contexts.

The finding that the video stimuli were the primary determinants of LR3PMS counts also carries significant theoretical implications. Contextual factors, such as the specific scenarios depicted in the video stimuli, may be more critical in eliciting mental state talk than previously thought (W. S. Hall et al., 1981; Meins et al., 2014). This observation complicates the assumption that cultural or linguistic differences alone account for variability in LR3PMS, pointing instead to the importance of the situational demands placed on speakers. Speakers across distinct cultural-linguistic contexts may differ in the frequency of their mental state talk not because of inherent cultural tendencies or features of the languages they speak, but because of variation in the frequency with which they encounter circumstances that require such

talk. Evidence of as much has been observed previously, albeit within a single culture (Frederickx & Hofmans, 2014).

Finally, the findings presented in Chapter 5 highlight the need to reconsider the role of talkativeness, and perhaps the role of language more broadly, in theories of mindreading. The significant relationship between Total Words Uttered and RMET performance suggests that aspects of linguistic behavior beyond the production of specific mental state terms may contribute to mindreading ability (K. Milligan et al., 2007). This finding suggests the possibility of a more general sociocognitive orientation or interactional style that is itself related to mindreading, expanding upon the narrow accounts linking language to mindreading traditionally emphasized in the literature (Gonzales et al., 2010; Scharfstein et al., 2011). The negative interaction between Total Words Uttered and LR3PMS further complicates this relationship, suggesting that the influence of mental state talk on mindreading varies as a function of overall speech production. These findings call for further research, as previous work has posited a singularly positive relationship between increased attention to the minds of others and mindreading ability. To the extent mental state talk indexes such attention, the fact it is related also to overall talkativeness across cultural-linguistic contexts is an intriguing one which requires additional investigation. It is possible, perhaps, that this reflects a universally human relationship between speech practices about the mind and their rootedness in individual variation in mindreading ability (Barron & Schneider, 2009; Sperber & Wilson, 2002).

Methodological Contributions

This dissertation also makes important methodological contributions to the study of mental state talk and mindreading. The novel approach of integrating a standardized set of video stimuli to elicit a comparable set of speech samples across three distinct cultural-linguistic contexts, and developing replicable methodologies by which to code these data for LR3PMS allowed for one of the first systematic and quantitative cross-linguistic comparisons of mental

state talk. This approach permitted the role of cultural-linguistic context and the role of the video stimuli themselves in driving participants' production of LR3PMS to be examined. The use of two distinct coding schemes - one conservative (Wellman and Estes Terms) and one broad (All Mental State Terms) - enabled a more nuanced analysis of mental state talk across languages. This dissertation thus served as a proof-of-concept of this method's utility and flexibility in coding mental state talk according to theoretically informed definitions of mental states.

Moreover, the decision to use the RMET as a measure of mindreading, despite its known limitations (Black, 2019; H. Kim et al., 2022), provided valuable insights into the cross-cultural applicability of this widely used tool. The results indicate that while the RMET was generally robust across the three cultural contexts studied, there were subtle differences in how participants from different cultures interpreted the stimuli, particularly when culturally specific coding schemes were applied. Though it is true that the application of such a tool must generally be done with caution, the findings reported here showed that even if scores were calculated in culturally specific ways, no significant differences in performance were observed across field sites. As such, there may be at least some sets of cultures wherein application of a flawed tool like the RMET may nevertheless be feasible. Nevertheless, future research should explore culturally-adapted versions of this test and others like it to better capture the nuances of mindreading across cultures. Even with the amendments made to this test, it still suffers from theoretical and design flaws (Adams Jr et al., 2010).

Remaining Questions and Future Directions

While this dissertation has addressed several key gaps in the literature, it also raises new questions that warrant further investigation. One of the most pressing questions is the extent to which the findings can be generalized beyond the three linguistic and cultural contexts studied. While the use of three unrelated languages in three distinct cultures provides a strong foundation for such comparison, it remains unclear whether the patterns observed here hold

across other languages and cultures. Given the emphasis on *lexical* references to third-party mental states, it is possible that running this study in languages which rely on indirect implication of comparable semantic content through phrasal metaphors might fail to replicate the results reported here (Lakoff, 2008; Winner, 1997). Additionally, highly polysynthetic and agglutinative languages may exhibit patterns of lexical reference that complicate these results (Chahuneau et al., n.d.; Passban, 2017; Proost, 2007). Indeed, future research would be well served to focus on encoding morphological units, the semantic content of which constitutes mental state reference, to avoid these potential pitfalls. In this same vein, future research should expand the sample to include a wider range of languages, particularly those from less-studied linguistic families, in order to determine whether the relationship between mental state talk and mindreading is truly universal or more context-dependent than previously thought.

Another important avenue for future research is the exploration of the developmental trajectory of the relationship between mental state talk and mindreading. While this dissertation focused on adult participants, the findings suggest that the relationship between language and mindreading may evolve over time and could differ significantly at various stages of development. Indeed, my findings contrast with some reported recently in the literature pertaining to the ontogenetic timeline over which the relationship between mental state talk and mindreading holds (Carr et al., 2018; K. Milligan et al., 2007). As such, additional longitudinal studies that track the development of mental state talk into adulthood could provide valuable insights into how these abilities interact and influence each other over the lifespan. Additionally, the role of speech quantity in mindreading performance, as highlighted in Chapter 5, suggests that future studies should explore the cognitive and social factors that contribute to overall speech production. Understanding why more talkative individuals perform better on mindreading tasks could reveal important aspects of social cognition that are currently overlooked. For example, it would be valuable to investigate whether this relationship is mediated by factors

such as social anxiety, extraversion, or general cognitive ability (Bagg et al., 2024; Nilsson & de López, 2016; Scharfstein et al., 2011). The dissertation's findings on the influence of video stimuli on LR3PMS production also open up new directions for research. Future studies should explore how different types of stimuli, including those that vary in emotional intensity, social complexity, or cultural relevance, might differentially elicit mental state talk across diverse populations, factors for which there is already some evidence (Parrigon et al., 2017). This line of inquiry could help clarify whether the observed patterns of mental state talk are attributable to the design of the stimuli I employed. As these were not rigorously psychometrically controlled in their design, it is possible that the results I observed are artifacts of such flaws.

Finally, this dissertation's findings raise important questions about the functional role of mental state talk in social interactions. While the data suggest a link between LR3PMS and mindreading, it remains unclear whether this relationship is causal or simply correlational. Experimental studies that manipulate the amount and type of mental state talk to which participants are exposed and then measure subsequent changes in mindreading performance, could help clarify the directionality of this relationship. Though challenging to implement, such studies would be particularly valuable in understanding the potential for interventions aimed at improving social cognition through targeted linguistic training (Durrleman et al., 2019). Current training studies have shown promising results among both disabled and typically-developing participant pools, though it is unclear over what duration their effects last and to what extent they may improve mindreading performance. Nevertheless, a growing body of research has shown that neurotypical participants' exposure to and engagement with written narrative is related to mindreading ability. As such, additional longitudinal and cross-cultural experiments in this vein may elucidate the functional role of mental state talk.

Conclusion

In conclusion, the current work provides a comprehensive examination of the relationship between mental state talk and mindreading ability across diverse linguistic and cultural contexts. By systematically quantifying LR3PMS and exploring their correlation with performance on the RMET, this dissertation addressed key gaps in the literature and challenged assumptions that have guided previous work in this area. The findings reported here highlight the importance of considering both the specific content of mental state talk and the broader context of linguistic behavior when studying social cognition, while underscoring the need for a more nuanced understanding of the role language and culture play in shaping the ability to understand others' minds. While the research presented here offers valuable insights, it also points to the challenges associated with studying mental-state talk and mindreading across cultural and linguistic contexts. The choices researchers make pertaining to what a mental state is, how to collect speech samples, what constitutes a mental state reference, and what tool one ought to use to measure mindreading may constitute inflection points at which the ability to detect the relevant phenomena improves or weakens. Though the findings presented here are relatively robust, future research should continue to develop more sophisticated models that account for the diverse ways in which people across the world talk about and understand the minds of others. By doing so, a more complete understanding of how talk about the mind varies and how it is related to mindreading, independent of cultural-linguistic context, may emerge.

Table 1

Pseudorandomized Orders of Video Stimuli Presentation

Order	Block 1					Block 2			
1	DO	DA	CO	PG	FB	SK	MG	IN	NV
2	MG	IN	SK	NV	FB	DA	CO	PG	DO
3	IN	PG	MG	CO	FB	SK	NV	DO	DA
4	SK	NV	DA	DO	FB	PG	IN	CO	MG
5	DO	DA	CO	PG	SK	MG	IN	NV	FB
6	MG	IN	SK	NV	DA	CO	PG	DO	FB
7	IN	PG	MG	CO	SK	NV	DO	DA	FB
8	SK	NV	DA	DO	PG	IN	CO	MG	FB

Note. An illustration of the pseudorandomized orders employed. CO = Cooperation, DA = Dangerous Animal, DO = Dominance, FB = False Belief, IN = Infidelity, MG = Mate Guarding, NV = Norm Violation, PG = Prestige, SK = Sickness.

Table 2

Variance Estimates and Standard Deviations of Variance Component Model 1 Random Effects

Variable	***Random Effect***	***Variance***	***Std.Dev.***	***Groups***
Participant ID nested within Field Site	Intercept	0.12428	0.3525	177
Interaction between Field Site and Video ID	Intercept	0.56473	0.7515	27
Video ID	Intercept	0.4947	0.7034	9
Field Site	Intercept	0.02472	0.1572	3
Number of observations: 1589				

Note. Variance estimates for each of the random effects included in the model vary substantially in the amount of variance attributed to each. Estimates are reported untransformed and thus represent the variance in the log of the counts of LR3PMS.

Table 3

Variance Estimates and Standard Deviations of Variance Component Model 3 Random Effects

Variable	***Random Effect***	***Variance***	***Std.Dev.***	***Groups***
Participant ID	Intercept	0.1238	0.3519	177
Video ID	Intercept	0.5296	0.7277	9
Field Site	Intercept	0.1103	0.3321	3
Number of observations: 1589				

Note. Variance estimates of the random effects included in the model vary substantially in the amount of attributed to each. Estimates are reported untransformed and thus represent the variance in the log of the counts of LR3PMS.

Table 4

Variance Estimates and Standard Deviations of Variance Component Model 1 Random Effects

Variable	***Random Effect***	***Variance***	***Std.Dev.***	***Groups***
Participant ID nested within Field Site	Intercept	0.03701	0.1924	177
Interaction between Field Site and Video ID	Intercept	0.05245	0.2290	27
Video ID	Intercept	0.0915	0.3025	9
Field Site	Intercept	0.0000000006461	0.0000254	3
Number of observations: 1589				

Note. Variance estimates for each of the random effects included in the model vary substantially in the amount of variance attributed to each. Estimates are reported untransformed and thus represent the variance in the log of the counts of LR3PMS.

Table 5

Variance Estimates and Standard Deviations of Variance Component Model 3 Random Effects

Variable	***Random Effect***	***Variance***	***Std.Dev.***	***Groups***
Participant ID	Intercept	0.0373	0.1932	177
Video ID	Intercept	0.0967	0.3110	9
Field Site	Intercept	0.0024	0.0486	3
Number of observations: 1589				

Note. Variance estimates of the random effects included in the model vary substantially in the amount of attributed to each. Estimates are reported untransformed and thus represent the variance in the log of the counts of LR3PMS.

Table 6

Correlations between four possible approaches to coding RMET Data

	Unstandardized Baron-Cohen	Standardized Baron-Cohen	Unstandardized Culturally Variable	Standardized Culturally Variable
Unstandardized Baron-Cohen	**1**	**0.9851788**	**0.9223612**	**0.9061009**
Standardized Baron-Cohen	**0.9851788**	**1**	**0.9032099**	**0.9191972**
Unstandardized Culturally Variable	**0.9223612**	**0.9032099**	**1**	**0.9820882**
Standardized Culturally Variable	**0.9061009**	**0.9191972**	**0.9820882**	**1**

Note. Each cell corresponds to the strength of the correlation coefficient between each of the four possible coding schemes for participant RMET responses. Even the weakest correlation indicated in the table above was found to be strongly positively correlated, $r(175) = .9032$, $p < .0001$. Additionally, a two-sample t-test was performed on the Unstandardized Baron-Cohen and Unstandardized Culturally Variable RMET scores to determine if there was sufficient evidence to suggest these scores were drawn from distributions whose means differed. Regardless of whether participant RMET scores had been coded using the Unstandardized Culturally Variable scheme ($M = 23.000$, $SD = 4.835$) or the Unstandardized Baron-Cohen scheme ($M = 22.232$, $SD = 5.077$), no difference was found between the sets of scores, $t(351.15) = 1.458$, $p = 0.146$.

Table 7

Correlations between two possible treatments of All Mental State Terms LR3PMS

	Count of LR3PMS	Rate of LR3PMS
Count of LR3PMS	1	0.1616074
Rate of LR3PMS	0.1616074	1

Note. Each cell corresponds to the strength of the correlation coefficient between each of the two possible ways of characterizing participant production of LR3PMS. Notably, the correlation between these two conceptions of LR3PMS is far weaker than the correlations between the different approaches to coding RMET performance. Consequently, it is possible that analyses using the *count* of LR3PMS might diverge significantly from analyses using the *rate* of LR3PMS.

Table 8

Correlations between two possible treatments of Wellman and Estes Terms LR3PMS

	Count of LR3PMS	Rate of LR3PMS
Count of LR3PMS	1	0.4600264
Rate of LR3PMS	0.4600264	1

Note. Each cell corresponds to the strength of the correlation coefficient between each of the two possible ways of characterizing participant production of LR3PMS. Notably, the correlation between these two conceptions of LR3PMS is far weaker than the correlations between the different approaches to coding RMET performance. Consequently, it is possible that analyses using the *count* of LR3PMS might diverge significantly from analyses using the *rate* of LR3PMS.

Table 9

Models Compared

Model	Specification
Model 1	RMET ~ Total Words Uttered
Model 2	RMET ~ LR3PMS Uttered
Model 3	RMET ~ Field Site
Model 4	RMET ~ Total Words Uttered + LR3PMS Uttered
Model 5	RMET ~ Total Words Uttered + Field Site
Model 6	RMET ~ LR3PMS Uttered + Field Site
Model 7	RMET ~ LR3PMS Uttered + Total Words Uttered + Field Site
Model 8	RMET ~ LR3PMS Uttered * Total Words Uttered
Model 9	RMET ~ Total Words Uttered * Field Site
Model 10	RMET ~ LR3PMS * Field Site
Model 11	RMET ~ Total Words Uttered * LR3PMS Uttered + Total Words Uttered * Field Site
Model 12	RMET ~ LR3PMS Uttered * Total Words Uttered + LR3PMS Uttered * Field Site
Model 13	RMET ~ Field Site * Total Words Uttered + Field Site * LR3PMS Uttered
Model 14	RMET ~ Total Words Uttered * LR3PMS Uttered + Total Words Uttered * Field Site + LR3PMS * Field Site
Model 15	RMET ~ Total Words Uttered * LR3PMS Uttered * Field Site

Note. All 15 models compared. RMET = Uncorrected Baron-Cohen Reading the Mind in the Eyes Test scores; Total Words Uttered = the total number of words produced by a given participant summed across all video stimuli descriptions; LR3PMS Uttered = the total number of lexical references to third-party mental states produced by a given participant summed across all video stimuli descriptions; Field Site = the field site from which a given participant was recruited, either the United States, China, or Morocco.

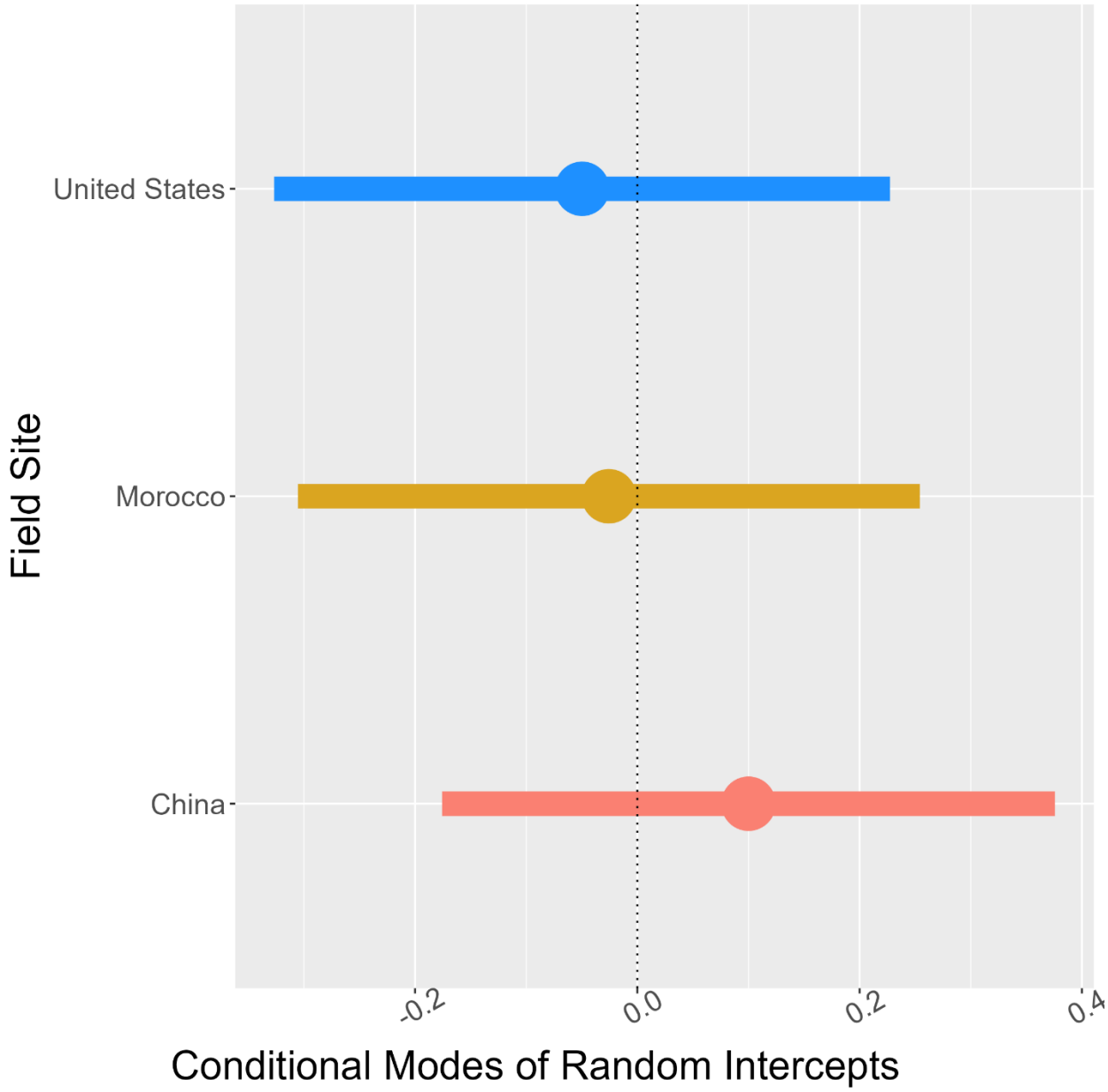
Figure 1

Stills Selected from Each of the Nine Video Stimuli

Note. A selection of stills from each of the video stimuli participants viewed. (A) “Cooperation” video stimulus. (B) “Dangerous Animal” video stimulus. (C) “Dominance” video stimulus. (D) “False Belief” video stimulus. (E) “Infidelity” video stimulus. (F) “Mate Guarding” video stimulus. (G) “Norm Violation” video stimulus. (H) “Prestige” video stimulus. (I) “Sickness” video stimulus.

Figure 2

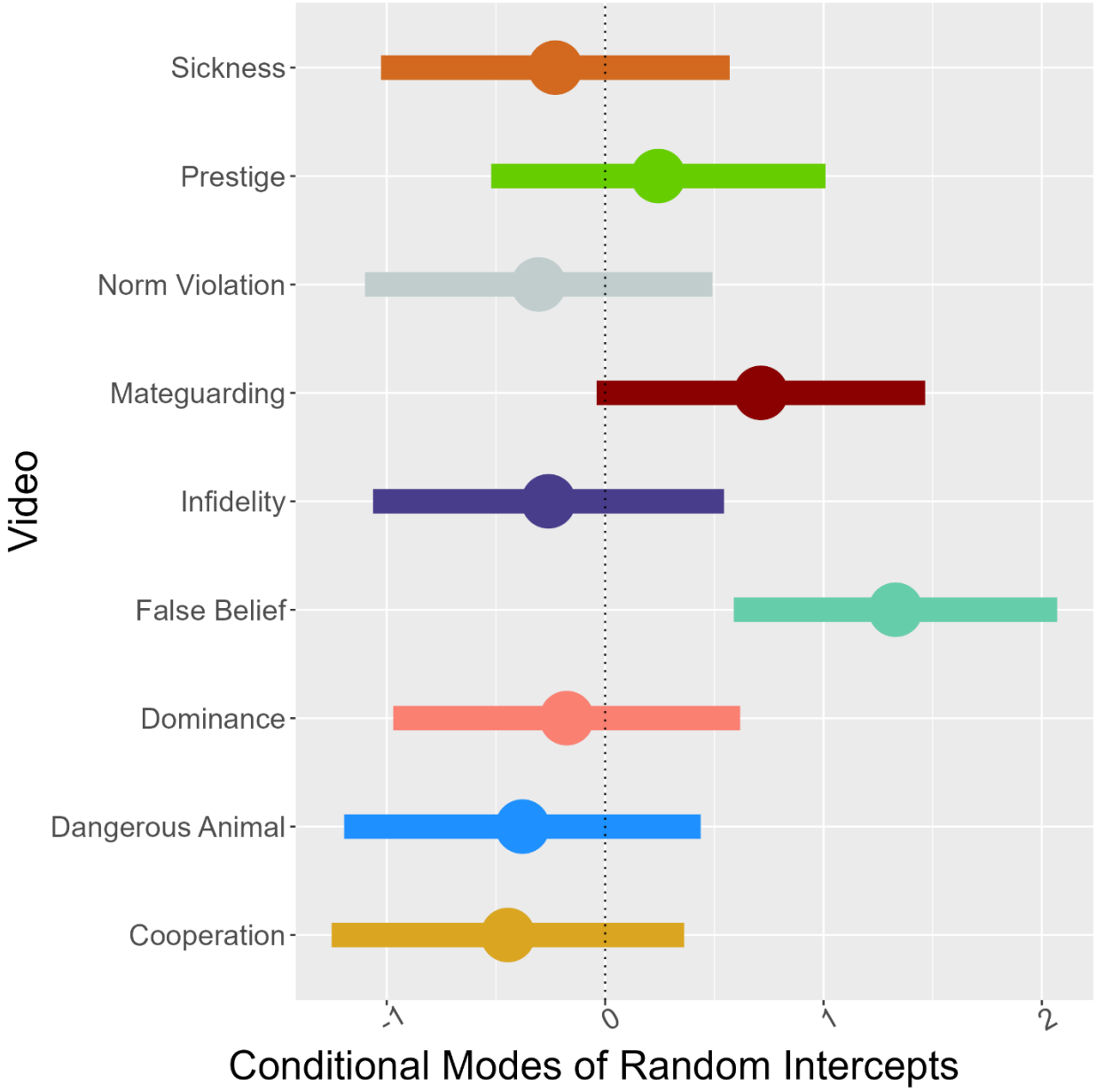
Conditional Modal Estimates of Random Intercepts for Levels of 'Field Site' Factor in VCM 1



Note. Conditional modal estimates of the random intercepts for each level of the 'Field Site' factor with 95% confidence intervals derived from VCM 1. Units are untransformed and therefore represent the difference between the log of the expected count of LR3PMS overall and the log of the expected count of LR3PMS for each level, holding all other variables constant.

Figure 3

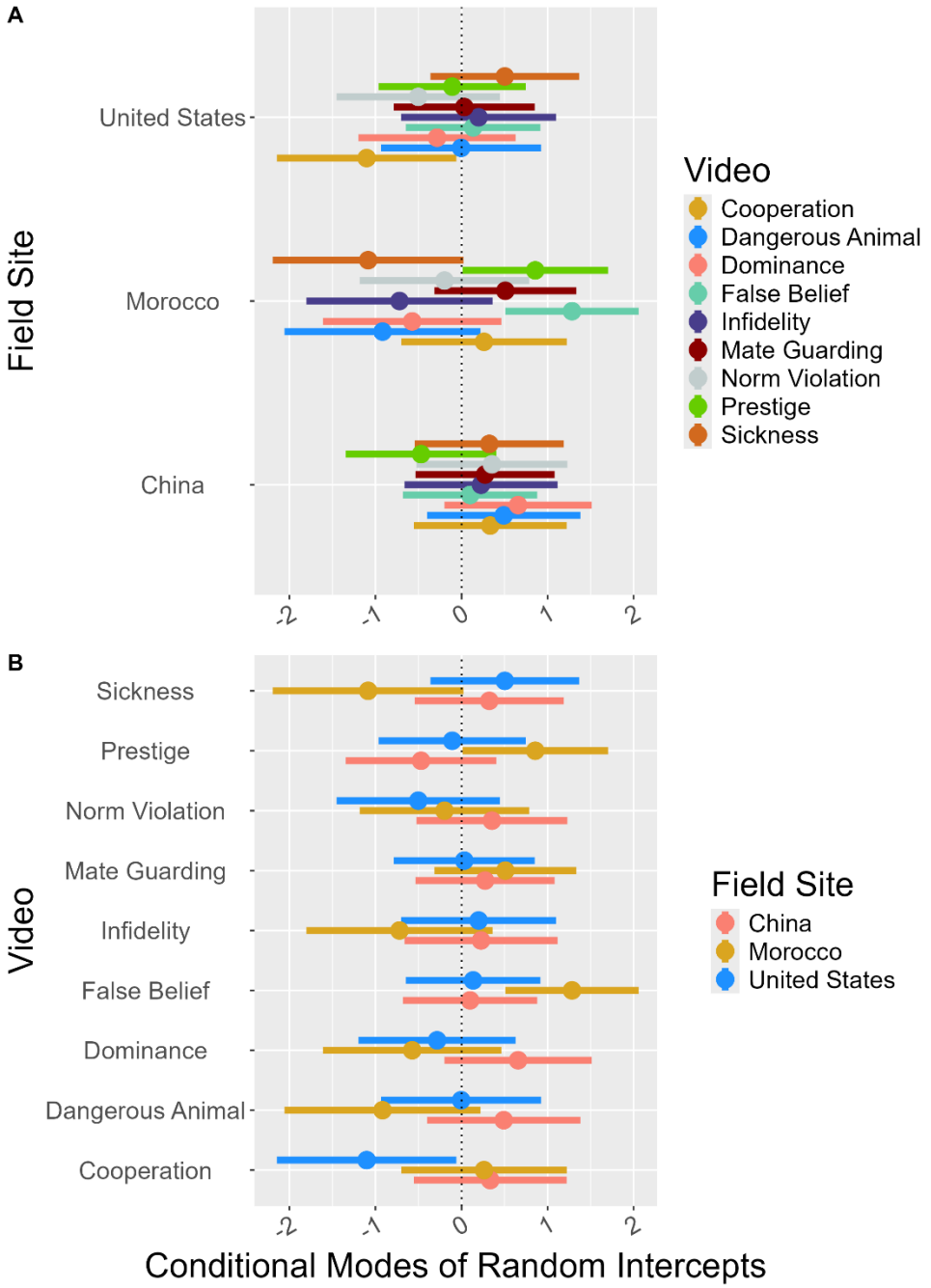
Conditional Modal Estimates of Random Intercepts for Levels of 'Video ID' Factor in VCM 1



Note. Conditional modal estimates of the random intercepts for each level of the 'Video ID' factor with 95% confidence intervals derived from VCM 1. Units are untransformed and therefore represent the difference between the log of the expected count of LR3PMS overall and the log of the expected count of LR3PMS for each level, holding all other variables constant.

Figure 4

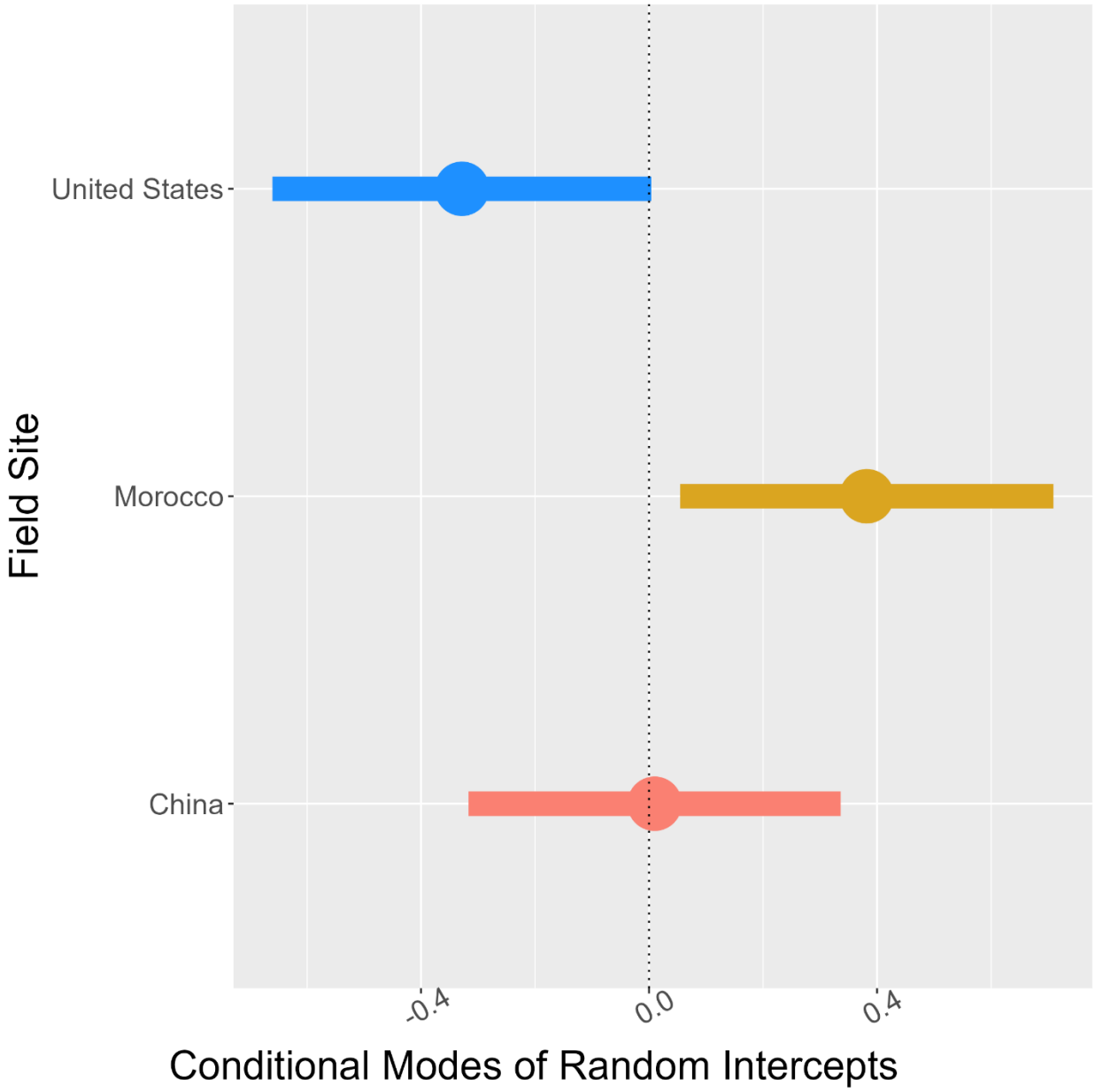
Conditional Modal Estimates of Random Intercepts for Levels of 'Video ID by Field Site' Interaction Factor in VCM 1



Note. Conditional modal estimates of the random intercepts for each level of the 'Video ID by Field Site' interaction with 95% confidence intervals from VCM 1. Units are untransformed and therefore represent the difference between the log of the expected count of LR3PMS overall and the log of the expected count of LR3PMS for each level, holding all other variables constant. (A) Estimates for each level of Video ID are grouped by Field Site on the Y axis. (B) Estimates for each level of Field Site are grouped by Video ID along the Y axis.

Figure 5

Conditional Modal Estimates of Random Intercepts for Levels of 'Field Site' Factor in VCM 3



Note. Conditional modal estimates of the random intercepts for each level of the 'Field Site' factor with 95% confidence intervals derived from VCM 3. Units are untransformed and therefore represent the difference between the log of the expected count of LR3PMS overall and the log of the expected count of LR3PMS for each level, holding all other variables constant.

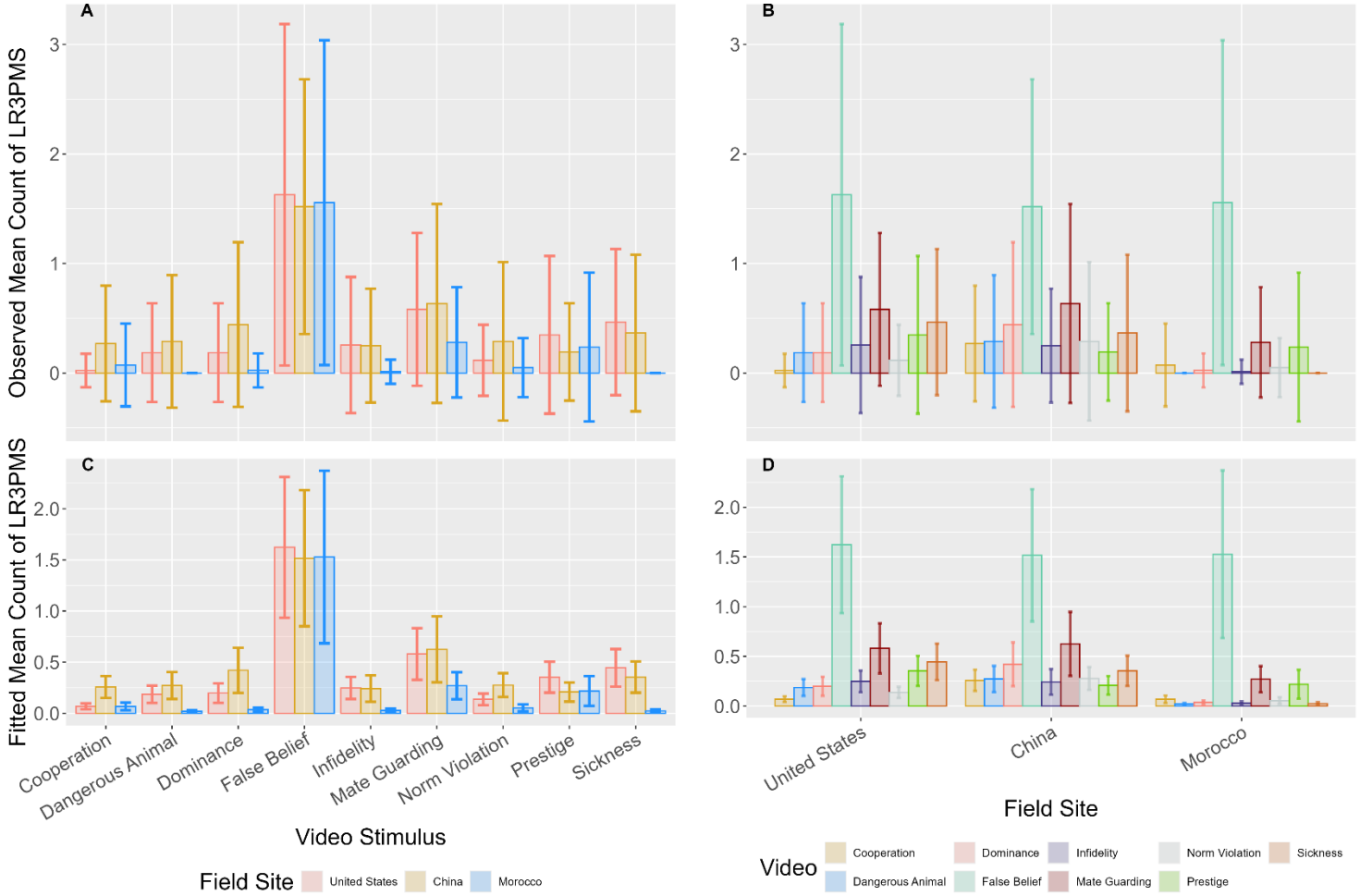
Figure 6

Conditional Modal Estimates of Random Intercepts for Levels of 'Video ID' Factor in VCM 3

Note. Conditional modal estimates of the random intercepts for each level of the 'Video ID' factor with 95% confidence intervals derived from VCM 3. Units are untransformed and therefore represent the difference between the log of the expected count of LR3PMS overall and the log of the expected count of LR3PMS for each level, holding all other variables constant.

Figure 7

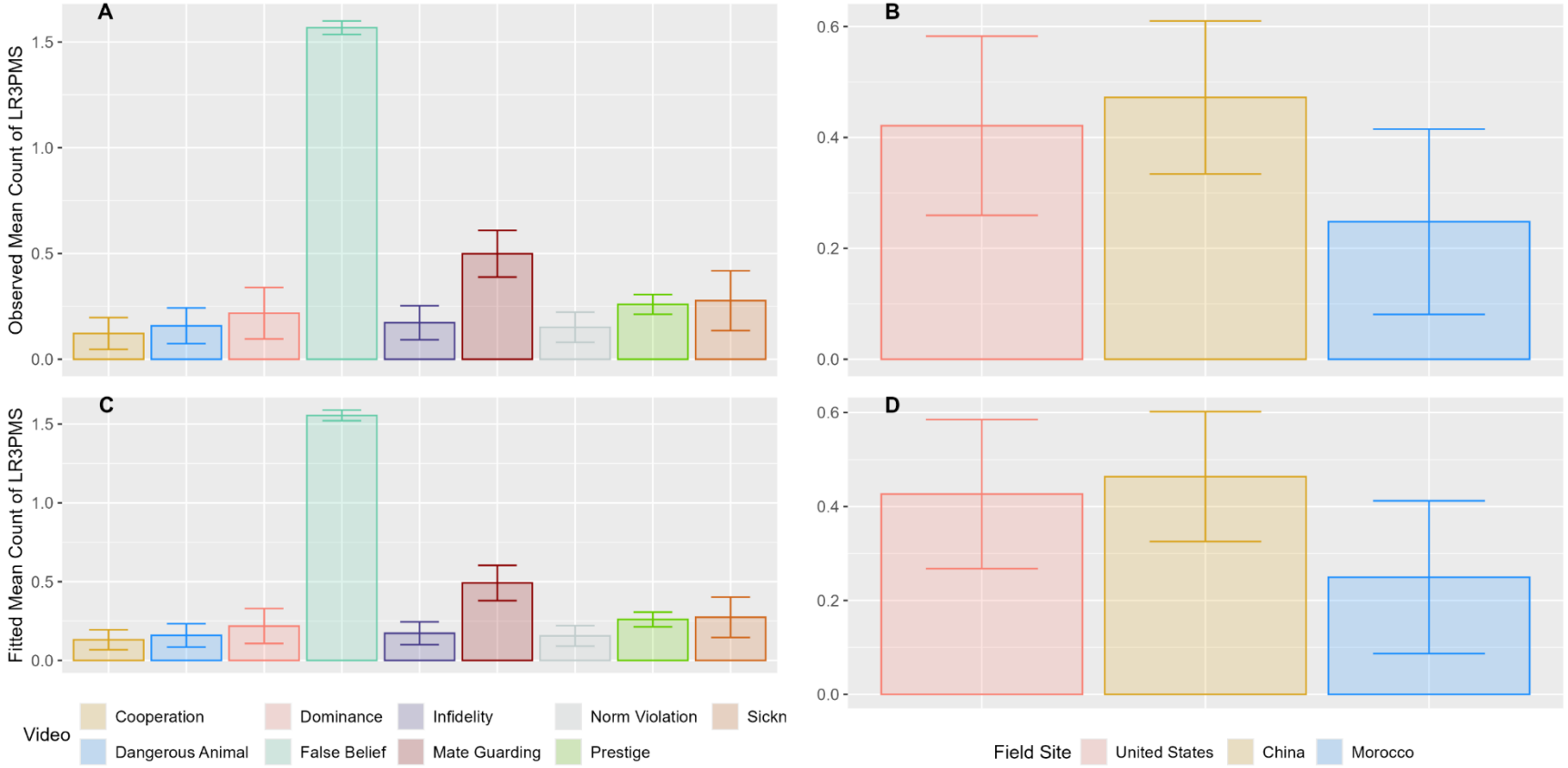
Observed and Fitted Estimates of Mean LR3PMS Counts for VCM1 Interaction Term



Note. VCM 1 observed and fitted estimates of the average count of LR3PMS. Error bars correspond to 95% confidence intervals. (A) Observed mean counts for each level of *Field Site* grouped by *Video ID* along the x-axis. (B) Observed mean counts for each level of *Video ID* grouped by *Field Site* along the x-axis. (C) Fitted estimates of mean counts for each level of *Field Site* grouped by *Video ID* along the x-axis. (D) Fitted estimates of mean counts for each level of *Video ID* grouped by *Field Site* along the x-axis.

Figure 8

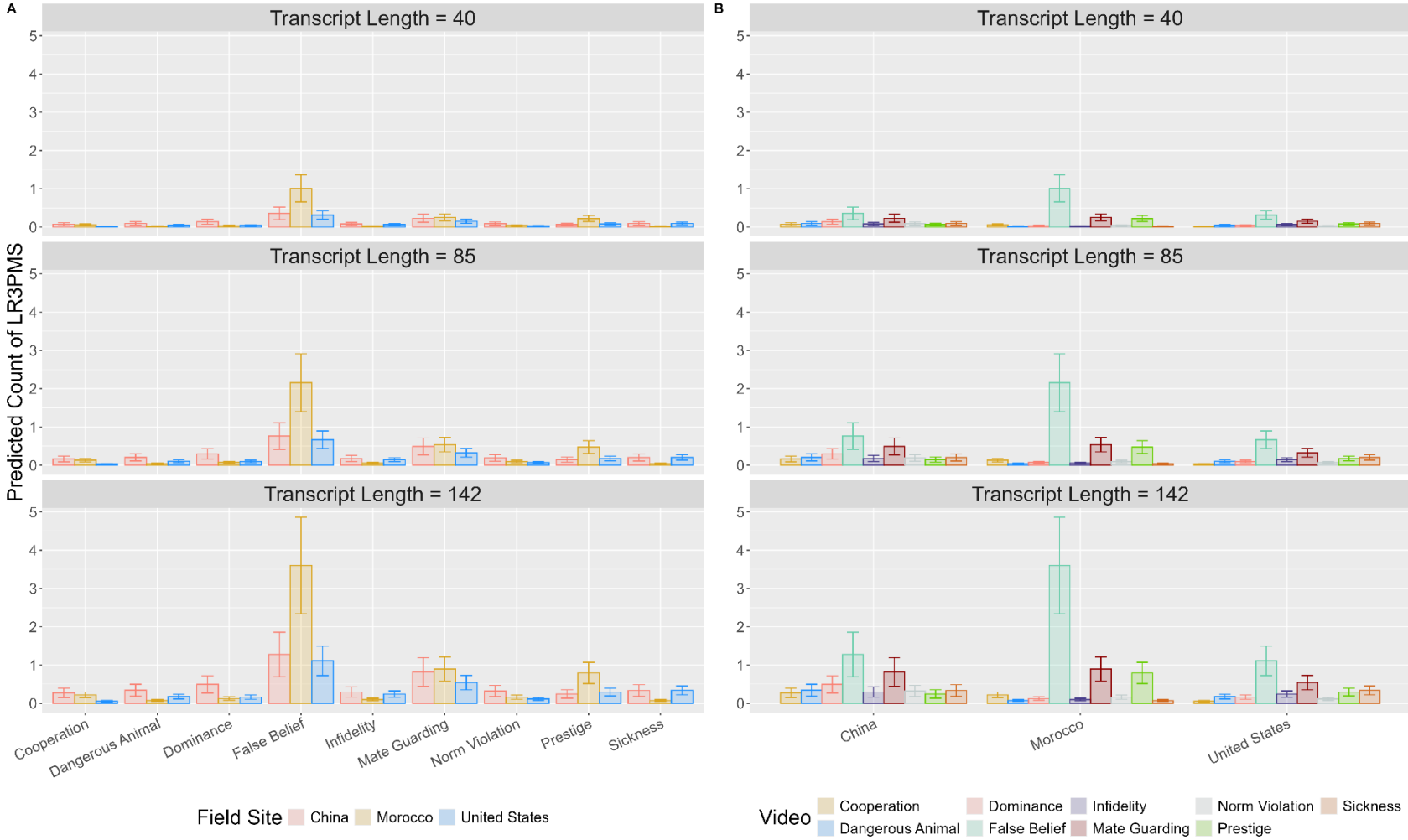
Observed and Fitted Estimates of Mean LR3PMS Counts for VCM1 Main Effects Term



Note. VCM 1 observed and fitted estimates of the average count of LR3PMS. Error bars correspond to 95% confidence intervals. (A) Observed mean count of LR3PMS organized by levels of *Video ID* factor. (B) Observed mean count of LR3PMS organized by levels of *Field Site* factor. (C) Fitted estimates of the average count of LR3PMS organized by levels of *Video ID* factor. (D) Fitted estimates of the average count of LR3PMS organized by levels of *Field Site* factor.

Figure 9

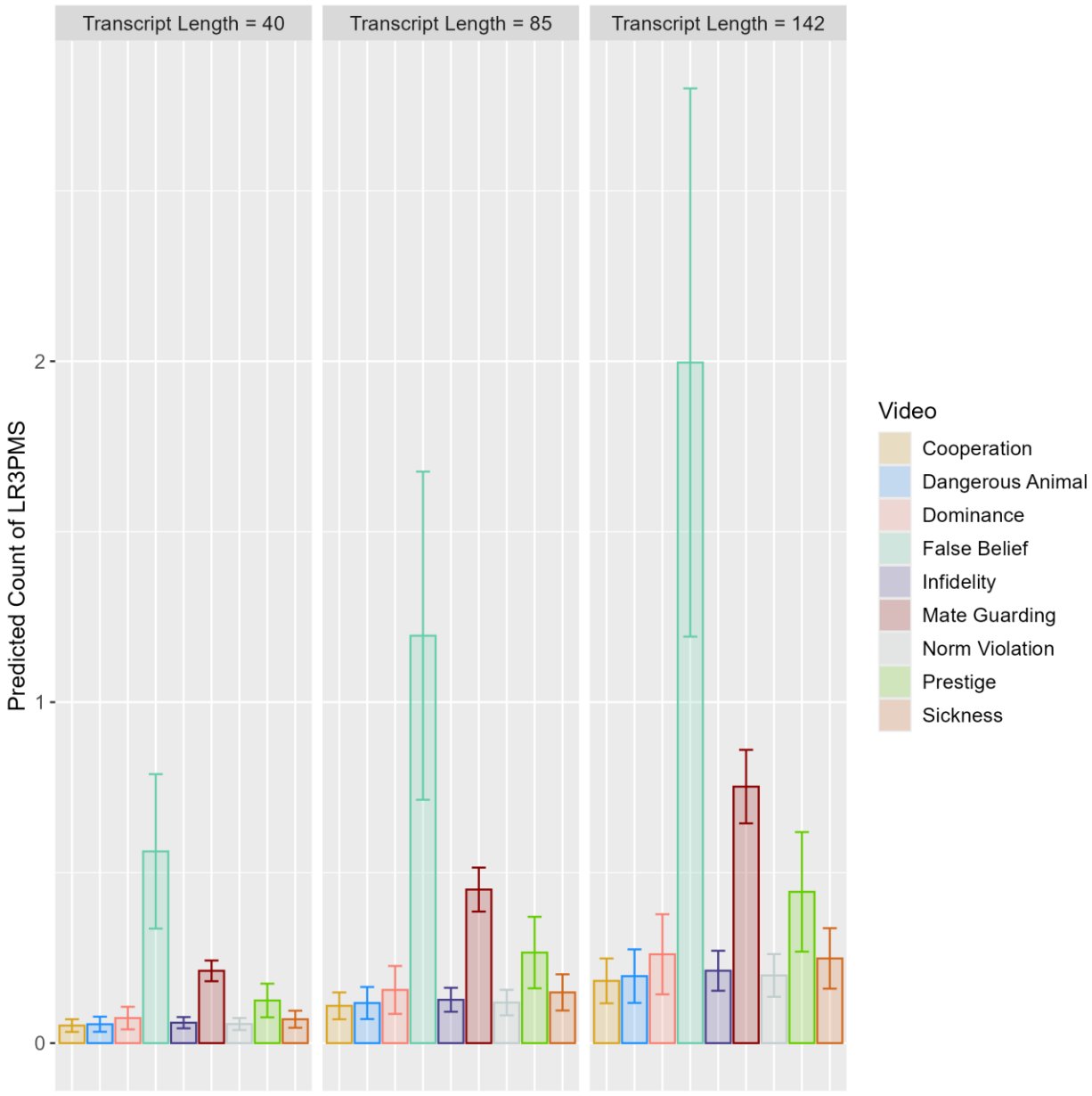
Predicted Estimates of Mean Counts of LR3PMS for the Lower Quartile, Median, and Upper Quartile Values of Transcript Length for VCM 1 Interaction Term



Note. VCM 1 predicted estimates of the average count of LR3PMS. Predictions were made holding all variables except transcript length constant. Selected transcript lengths corresponded to the lower quartile, median, and upper quartile values of the variable. Error bars correspond to the 95% confidence interval. (A) *Field Site* grouped by *Video ID*. (B) *Video ID* grouped by *Field Site.*

Figure 10

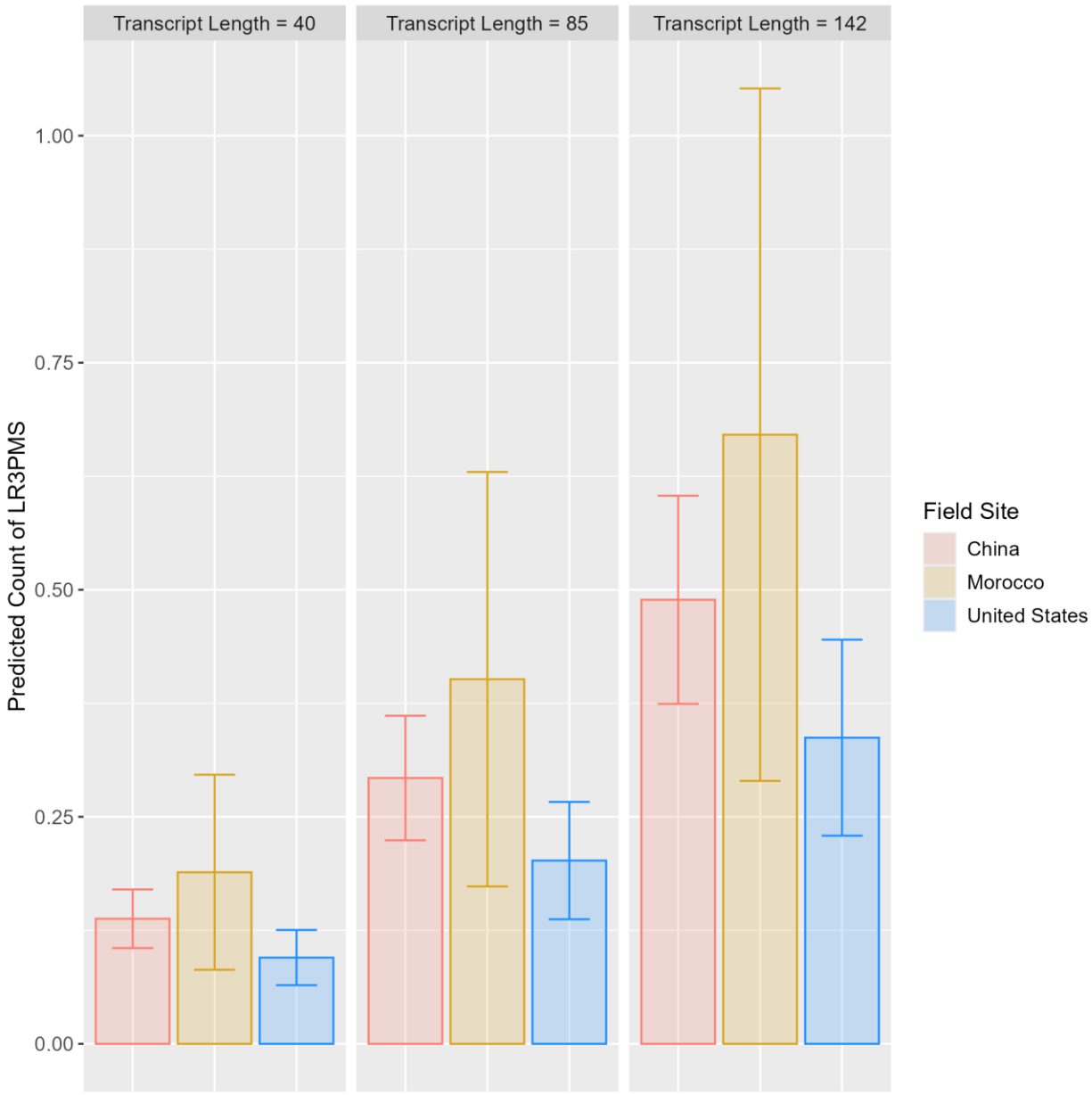
VCM 1 Predicted Estimates of Mean Counts of LR3PMS for the Lower Quartile, Median, and Upper Quartile Values of Transcript Length Across Each Level of 'Video ID'



Note. VCM 1 predicted estimates of the average count of LR3PMS organized by levels of *Video ID*. Predictions were made holding all variables except transcript length constant. Selected transcript lengths corresponded to the lower quartile, median, and upper quartile values of the variable. Error bars correspond to the 95% confidence interval.

Figure 11

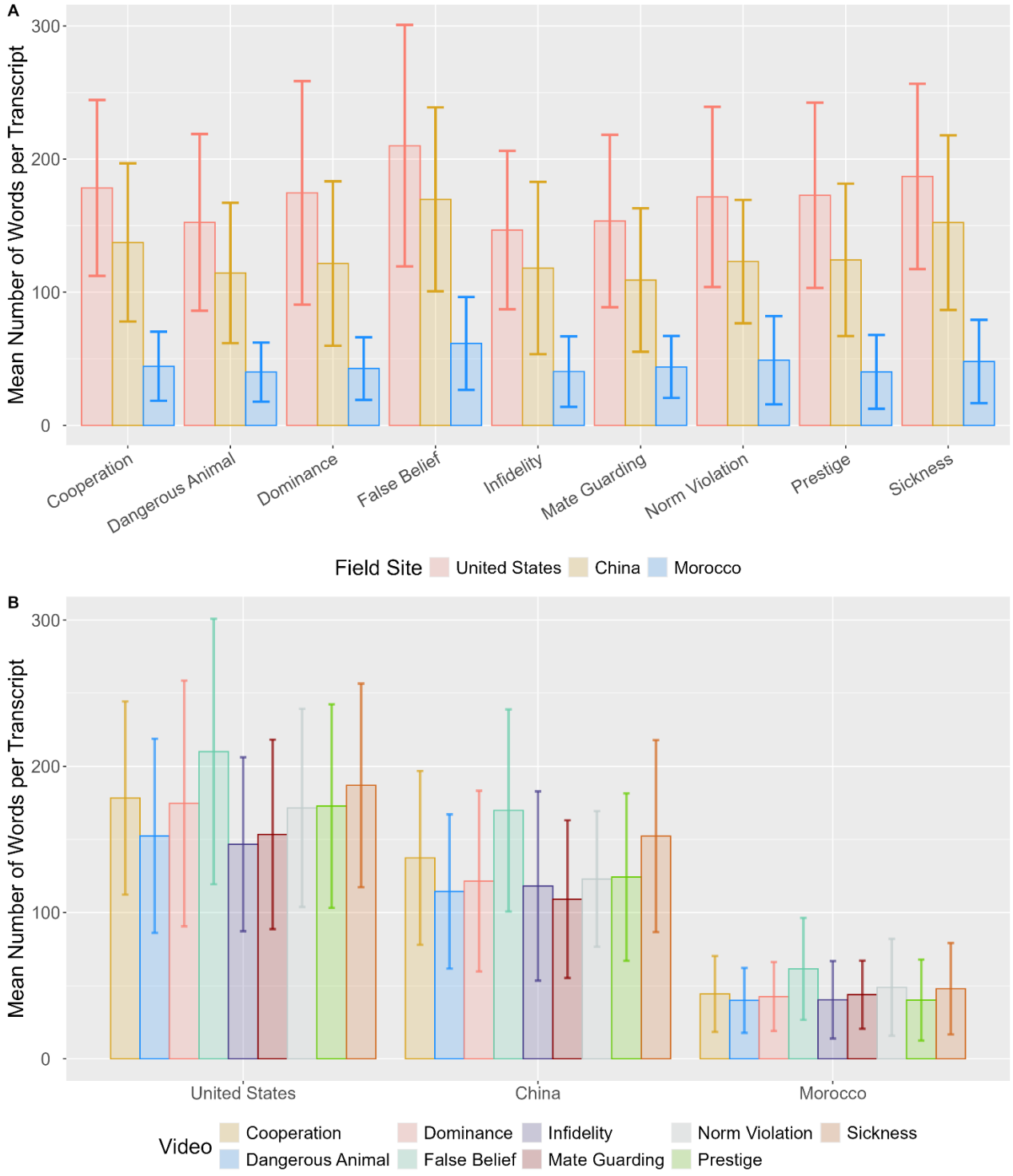
VCM 1 Predicted Estimates of Mean Counts of LR3PMS for the Lower Quartile, Median, and Upper Quartile Values of Transcript Length Across Each Level of 'Field Site'



Note. VCM 1 predicted estimates of the average count of LR3PMS organized by levels of *Field Site*. Predictions were made holding all variables except transcript length constant. Selected transcript lengths corresponded to the lower quartile, median, and upper quartile values of the variable. Error bars correspond to the 95% confidence interval.

Figure 12

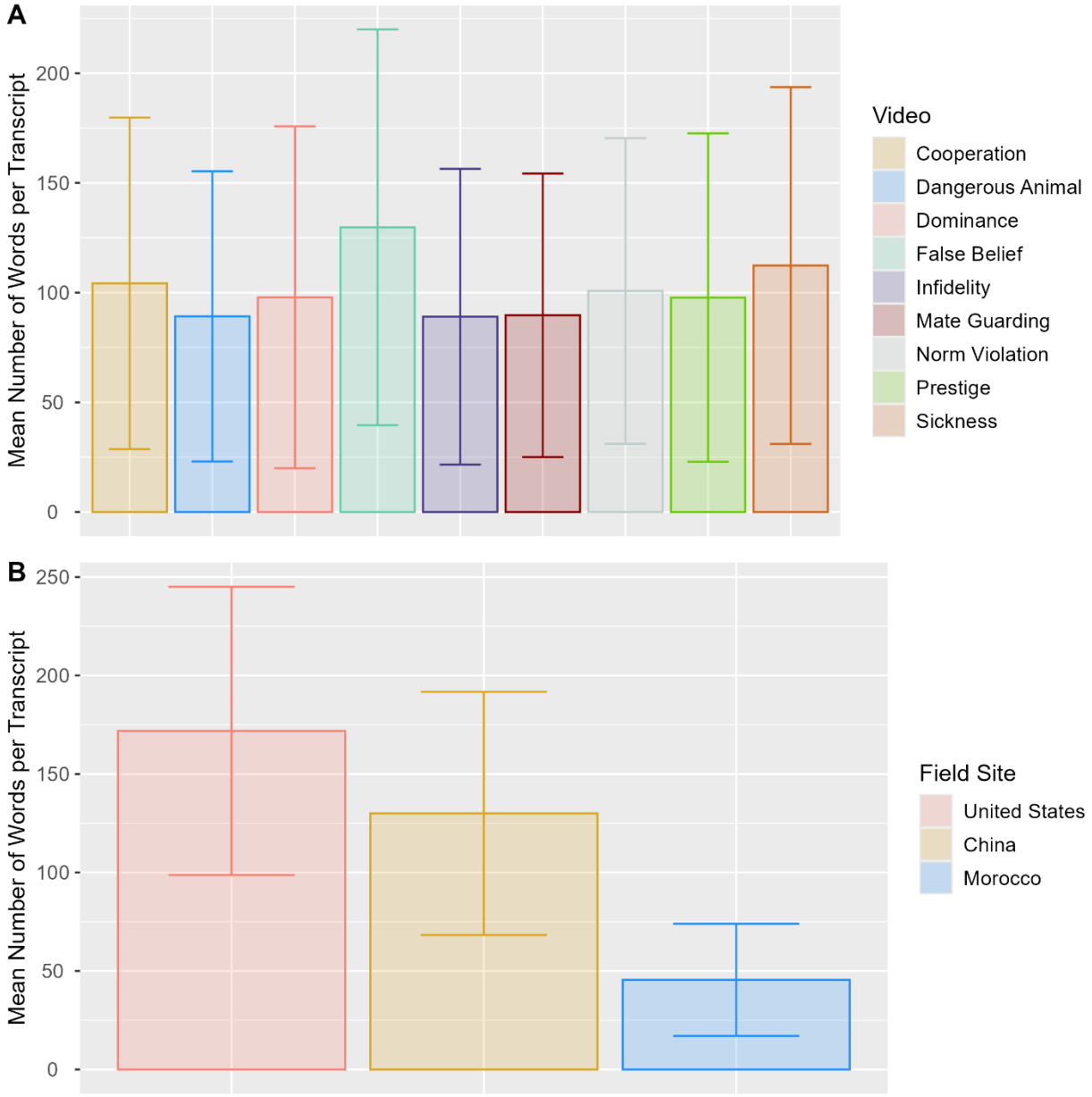
Observed Mean Transcript Length for Each Level of 'Video ID' Grouped by Each Level of 'Field Site'



Note. Mean observed transcript lengths. Error bars correspond to the 95% confidence interval. (A) *Field Site* grouped by *Video ID* along the x-axis. (B) *Video ID* grouped by *Field Site* along the x-axis.

Figure 13

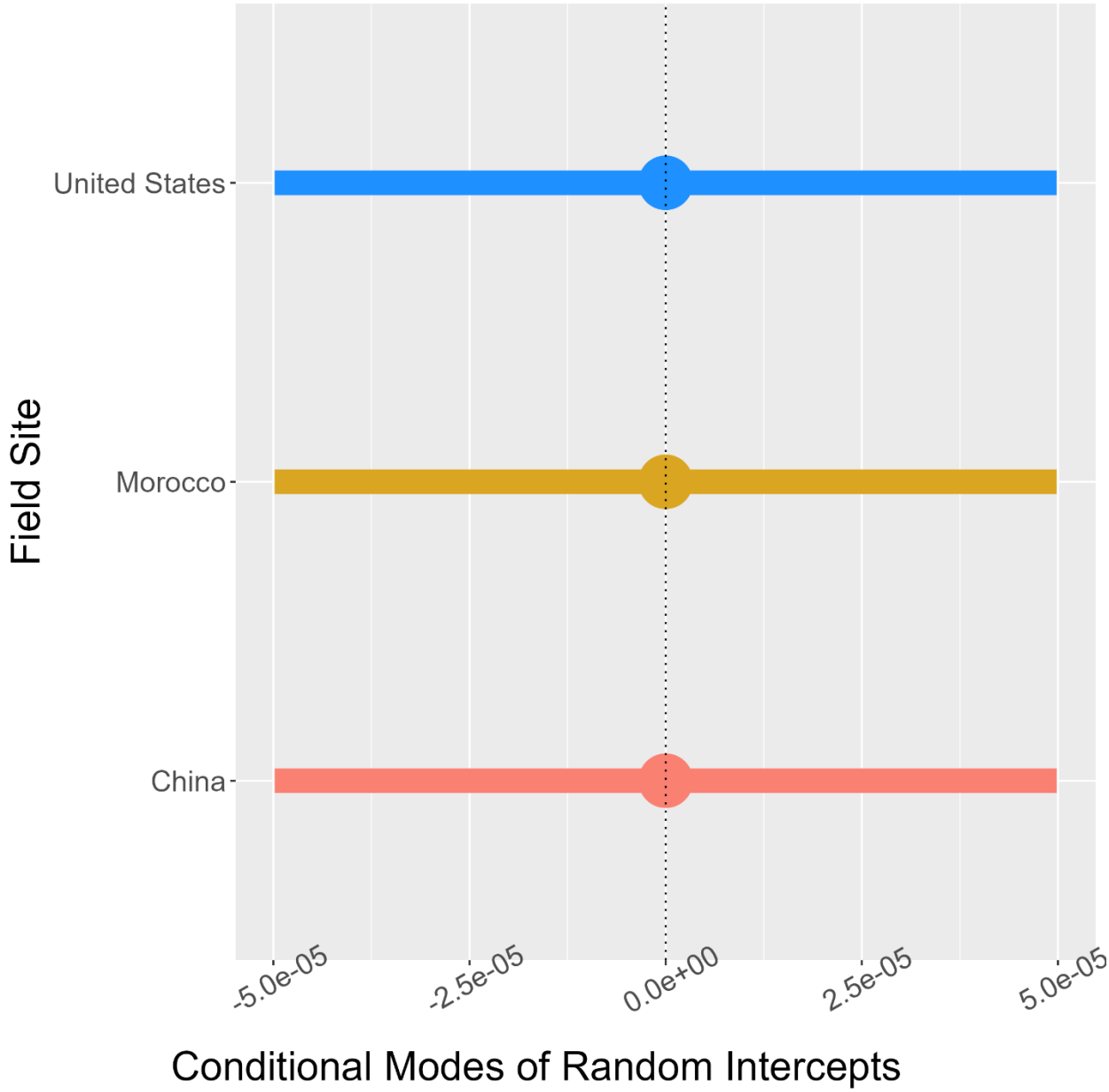
Observed Mean Transcript Length for Video ID and Field Site



Note. Mean observed transcript lengths. Error bars correspond to the 95% confidence interval. (A) Across levels of *Video ID* factor. (B) Across levels of *Field Site* factor.

Figure 14

Conditional Modal Estimates of Random Intercepts for Levels of 'Field Site' Factor in VCM 1



Note. Conditional modal estimates of the random intercepts for each level of the 'Field Site' factor with 95% confidence intervals derived from VCM 1. Units are untransformed and therefore represent the difference between the log of the expected count of LR3PMS overall and the log of the expected count of LR3PMS for each level, holding all other variables constant.

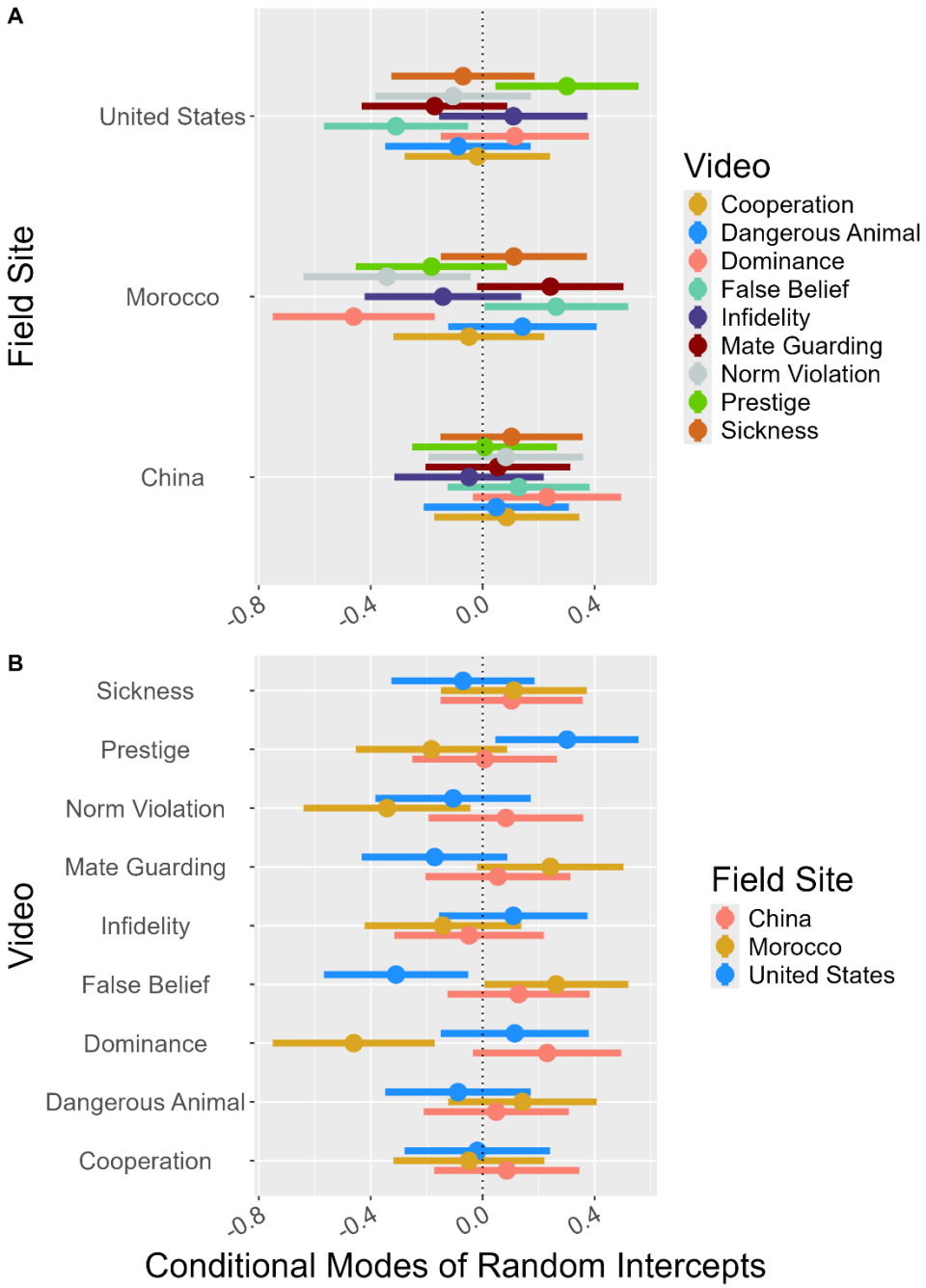
Figure 15

Conditional Modal Estimates of Random Intercepts for Levels of 'Video ID' Factor in VCM 1

Note. Conditional modal estimates of the random intercepts for each level of the 'Video ID' factor with 95% confidence intervals derived from VCM 1. Units are untransformed and therefore represent the difference between the log of the expected count of LR3PMS overall and the log of the expected count of LR3PMS for each level, holding all other variables constant.

Figure 16

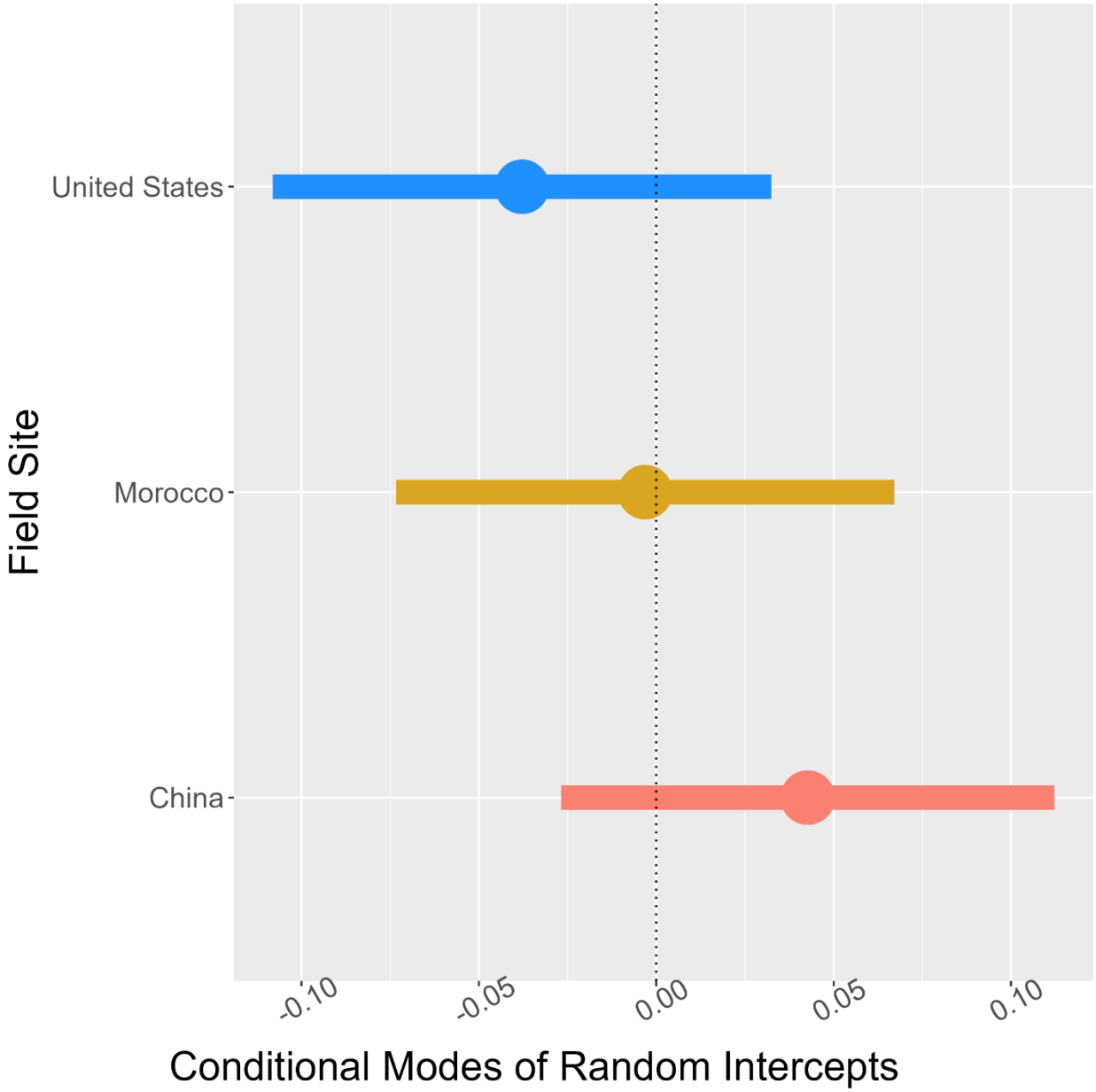
Conditional Modal Estimates of Random Intercepts for Levels of Interaction Factor in VCM 1



Note. Conditional modal estimates of the random intercepts for each level of the Interaction factor in VCM 1 with 95% confidence intervals. Units are untransformed and therefore represent the difference between the log of the expected count of LR3PMS overall and the log of the expected count of LR3PMS for each level, holding all other variables constant. (A) Estimates for each level of *Video ID* are grouped by levels of *Field Site* on the y-axis. (B) Estimates for each level of *Field Site* are grouped by levels of *Video ID* on the y-axis.

Figure 17

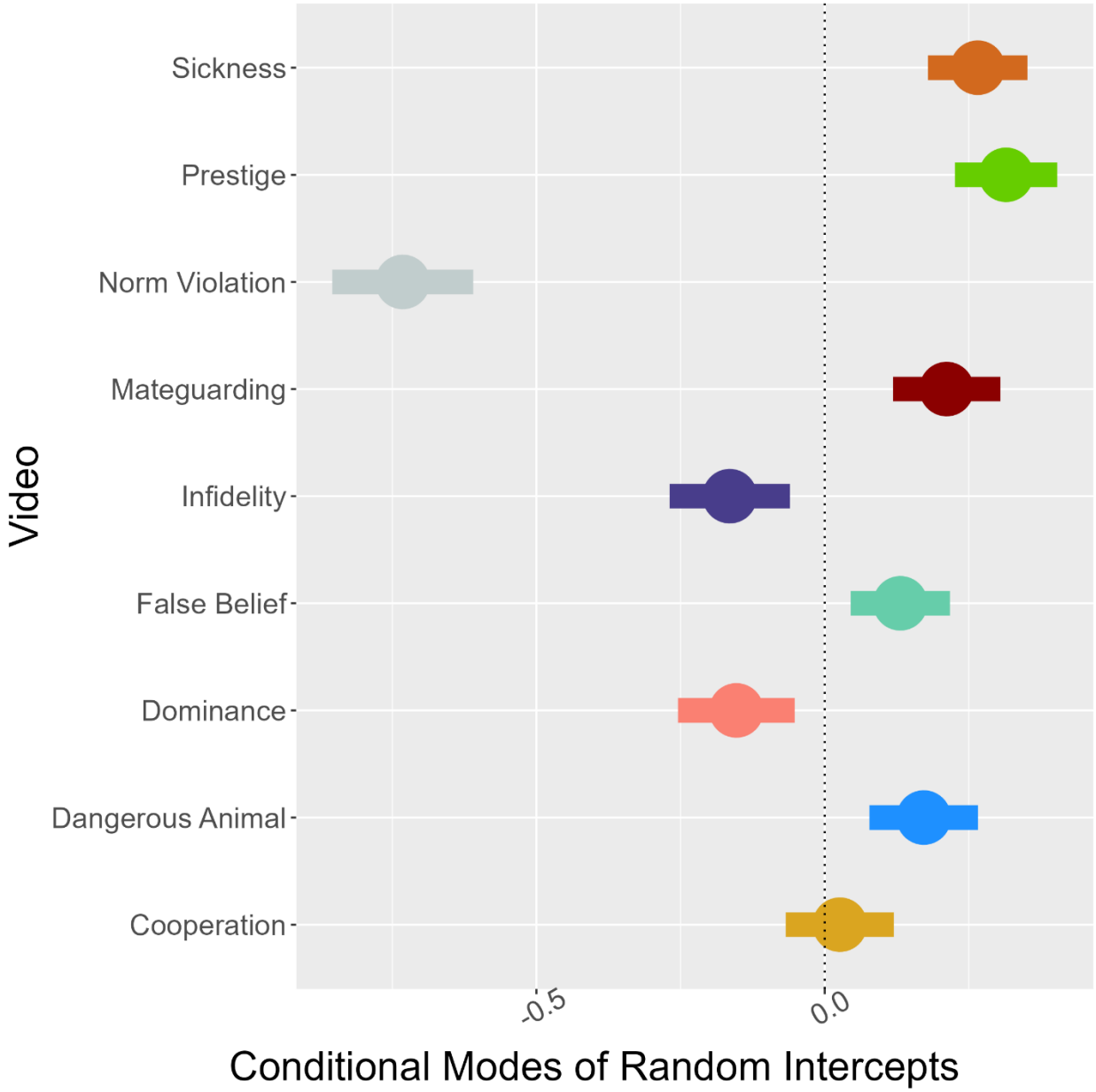
Conditional Modal Estimates of Random Intercepts for Levels of 'Field Site' Factor in VCM 3



Note. Conditional modal estimates of the random intercepts for each level of the 'Field Site' factor with 95% confidence intervals derived from VCM 3. Units are untransformed and therefore represent the difference between the log of the expected count of LR3PMS overall and the log of the expected count of LR3PMS for each level, holding all other variables constant.

Figure 18

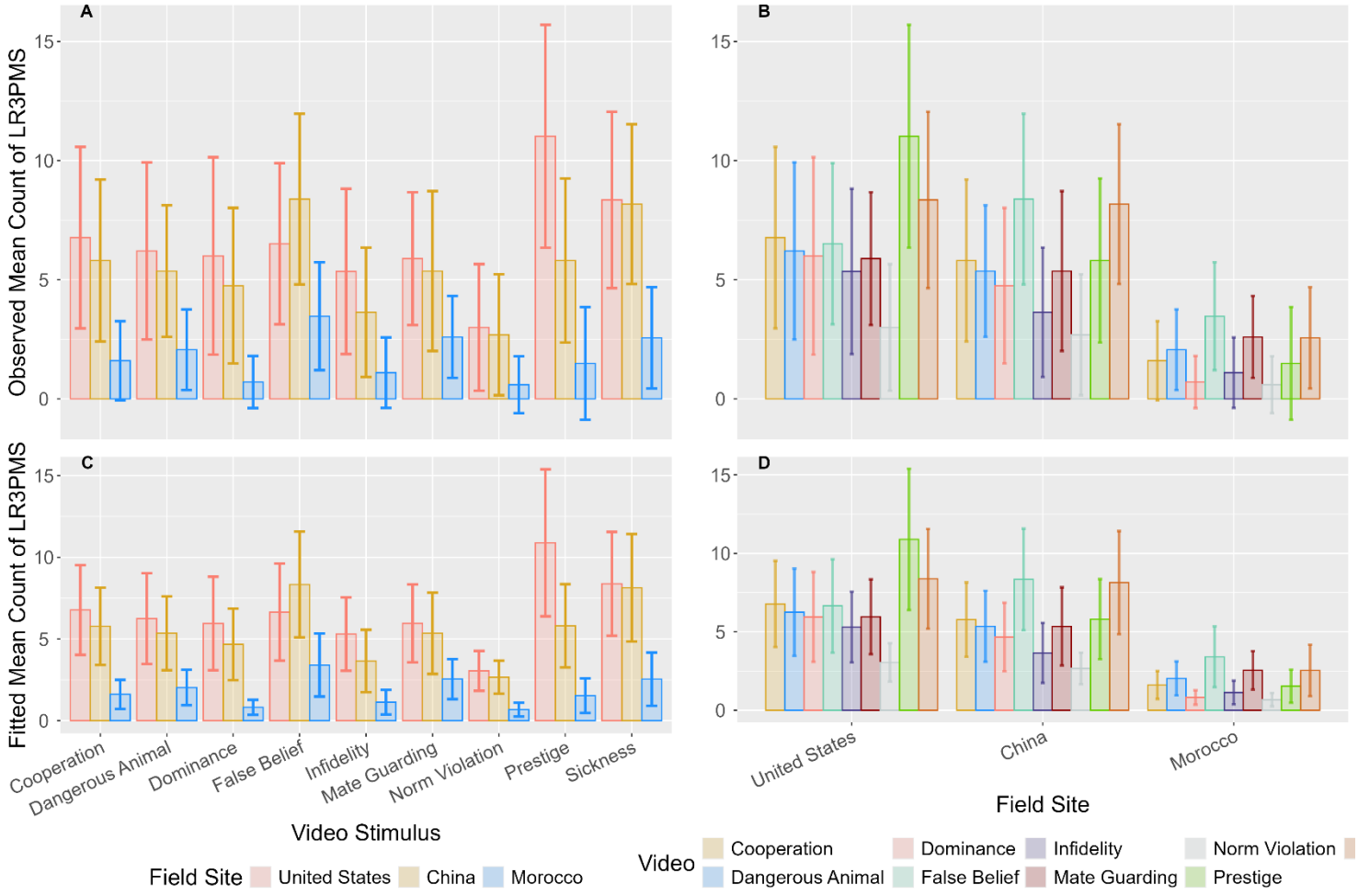
Conditional Modal Estimates of Random Intercepts for Levels of 'Video ID' Factor in VCM 3



Note. Conditional modal estimates of the random intercepts for each level of the 'Video ID' factor with 95% confidence intervals derived from VCM 3. Units are untransformed and therefore represent the difference between the log of the expected count of LR3PMS overall and the log of the expected count of LR3PMS for each level, holding all other variables constant.

Figure 19

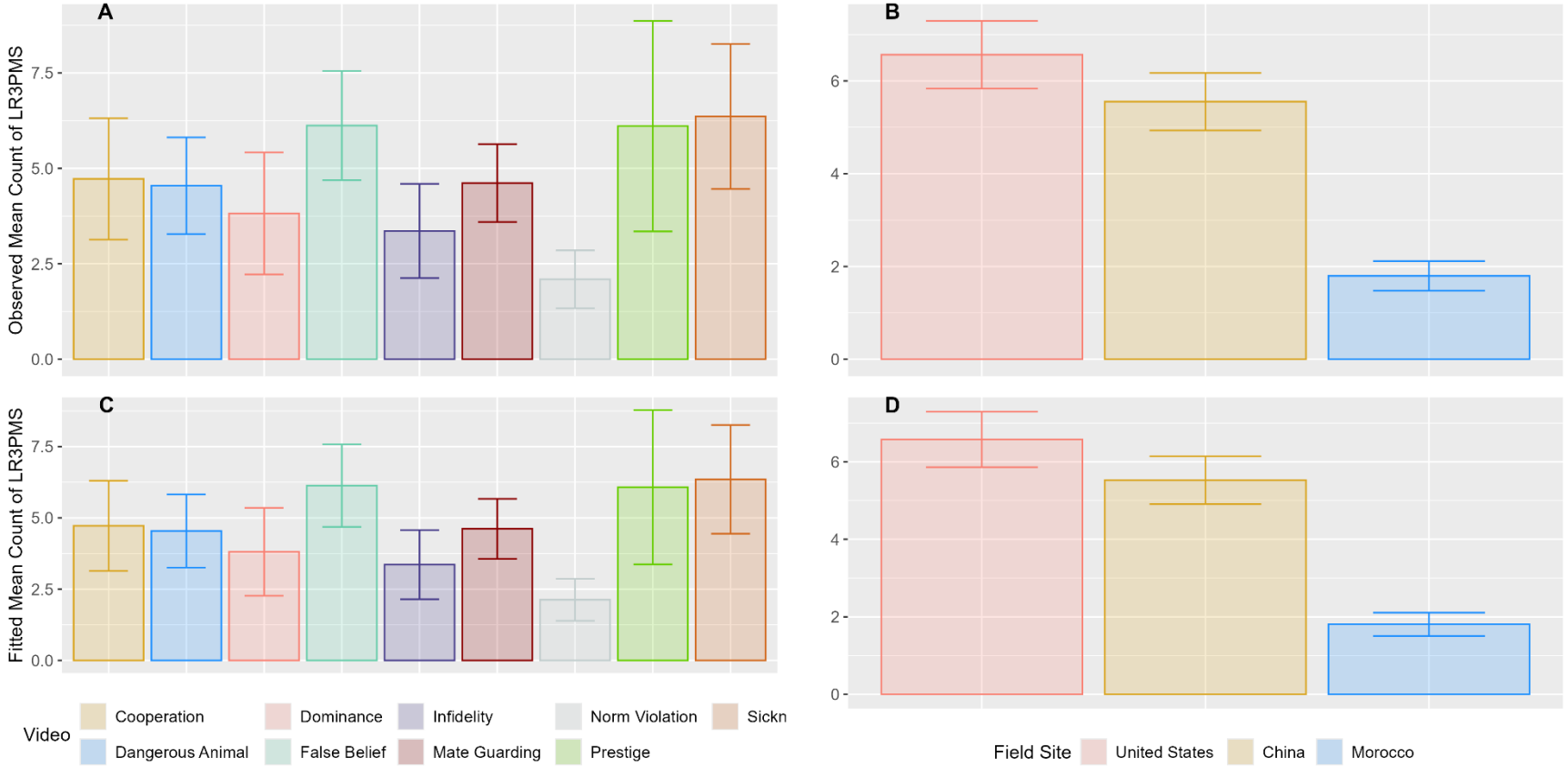
Observed and Fitted Estimates of Mean LR3PMS Counts for VCM1 Interaction Term



Note. VCM 1 observed and fitted estimates of the average count of LR3PMS. Error bars correspond to 95% confidence intervals. (A) Observed mean counts for each level of *Field Site* grouped by *Video ID* along the x-axis. (B) Observed mean counts for each level of *Video ID* grouped by *Field Site* along the x-axis. (C) Fitted estimates of mean counts for each level of *Field Site* grouped by *Video ID* along the x-axis. (D) Fitted estimates of mean counts for each level of *Video ID* grouped by *Field Site* along the x-axis.

Figure 20

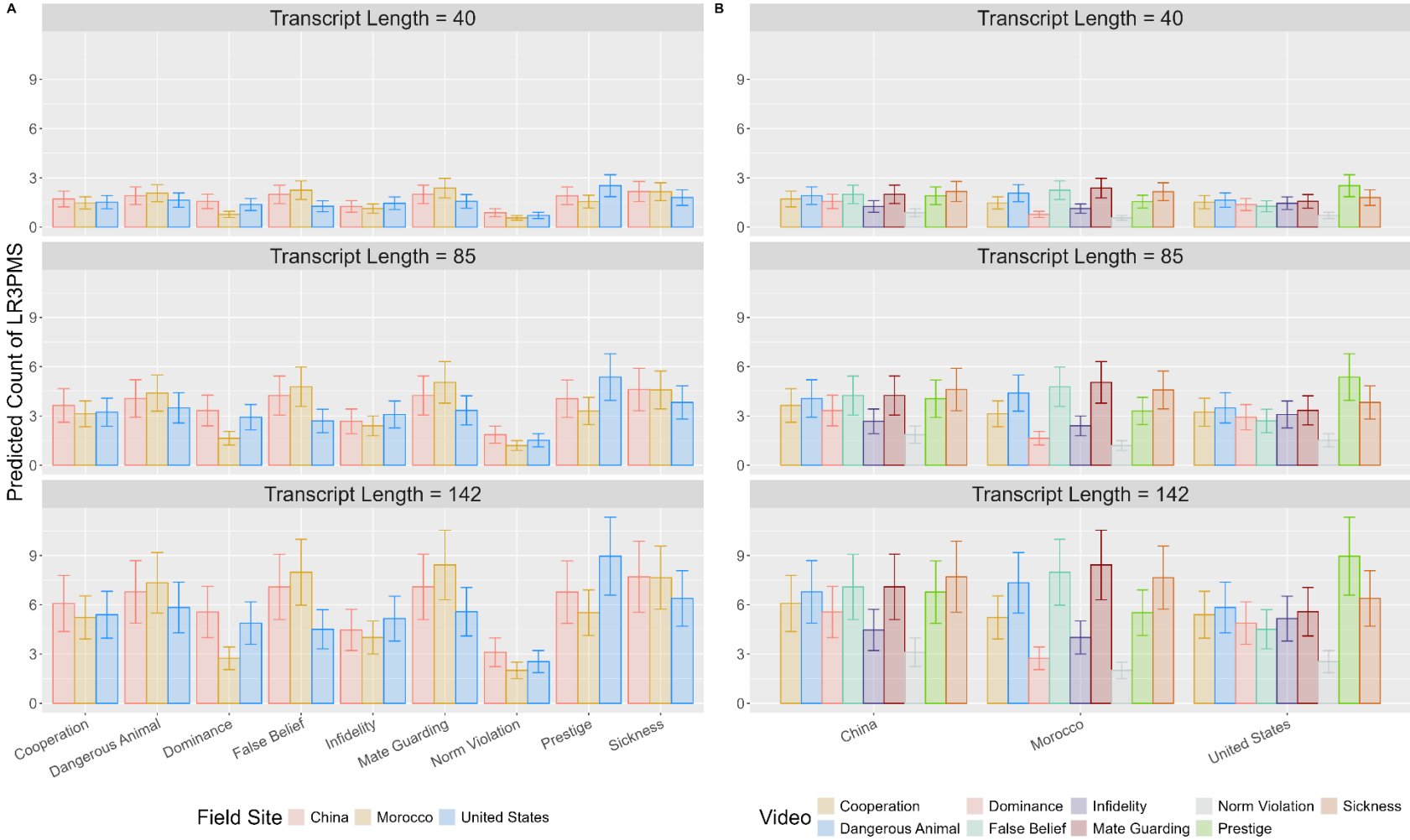
Observed and Fitted Estimates of Mean LR3PMS Counts for VCM1 Main Effects Term



Note. VCM 1 observed and fitted estimates of the average count of LR3PMS. Error bars correspond to 95% confidence intervals. (A) Observed mean count of LR3PMS organized by levels of *Video ID* factor. (B) Observed mean count of LR3PMS organized by levels of *Field Site* factor. (C) Fitted estimates of the average count of LR3PMS organized by levels of *Video ID* factor. (D) Fitted estimates of the average count of LR3PMS organized by levels of *Field Site* factor.

Figure 21

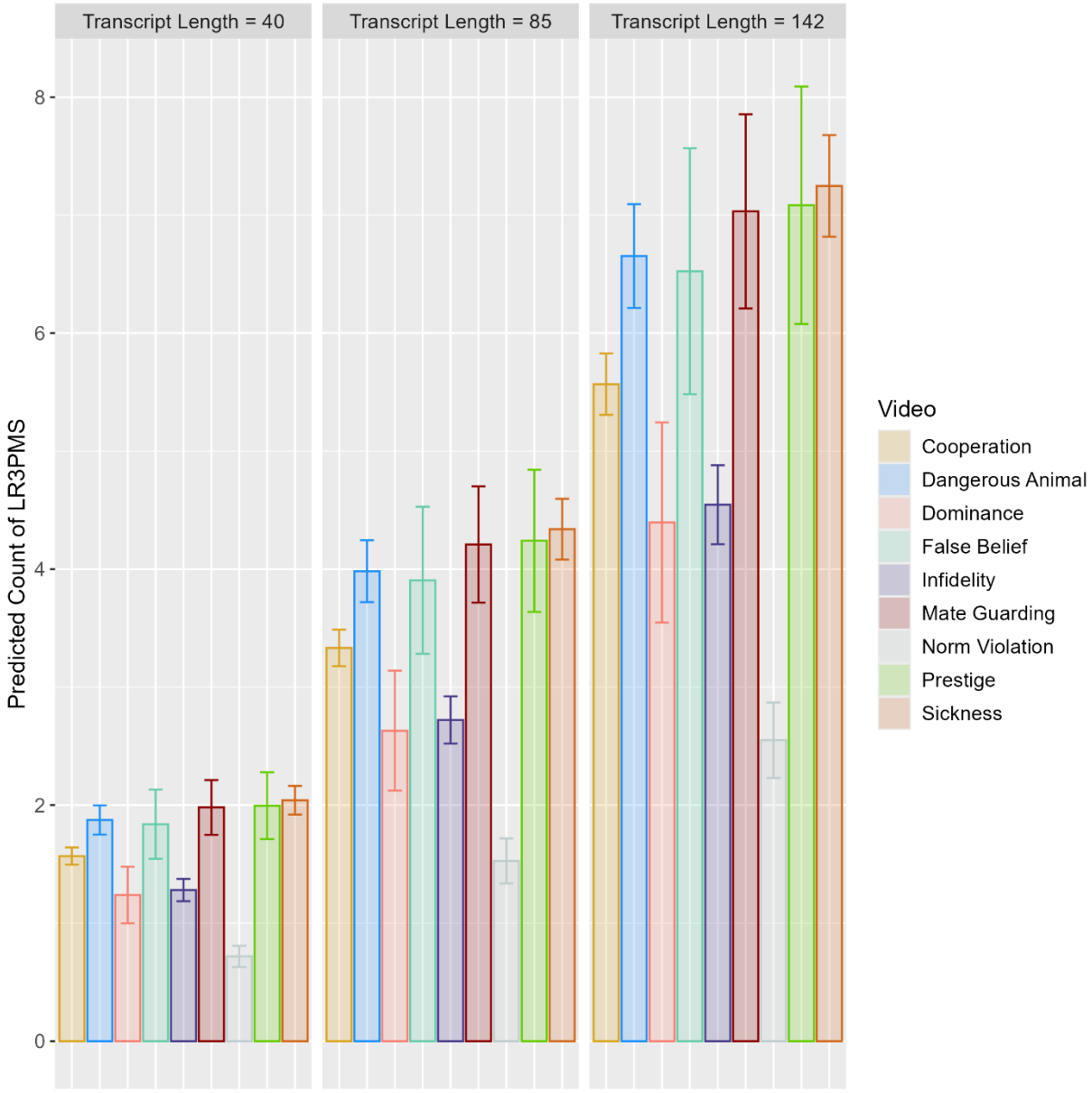
Predicted Estimates of Mean Counts of LR3PMS for the Lower Quartile, Median, and Upper Quartile Values of Transcript Length for VCM 1 Interaction Term



Note. VCM 1 predicted estimates of the average count of LR3PMS. Predictions were made holding all variables except transcript length constant. Selected transcript lengths corresponded to the lower quartile, median, and upper quartile values of the variable. Error bars correspond to the 95% confidence interval. (A) *Field Site* grouped by *Video ID*. (B) *Video ID* grouped by *Field Site.*

Figure 22

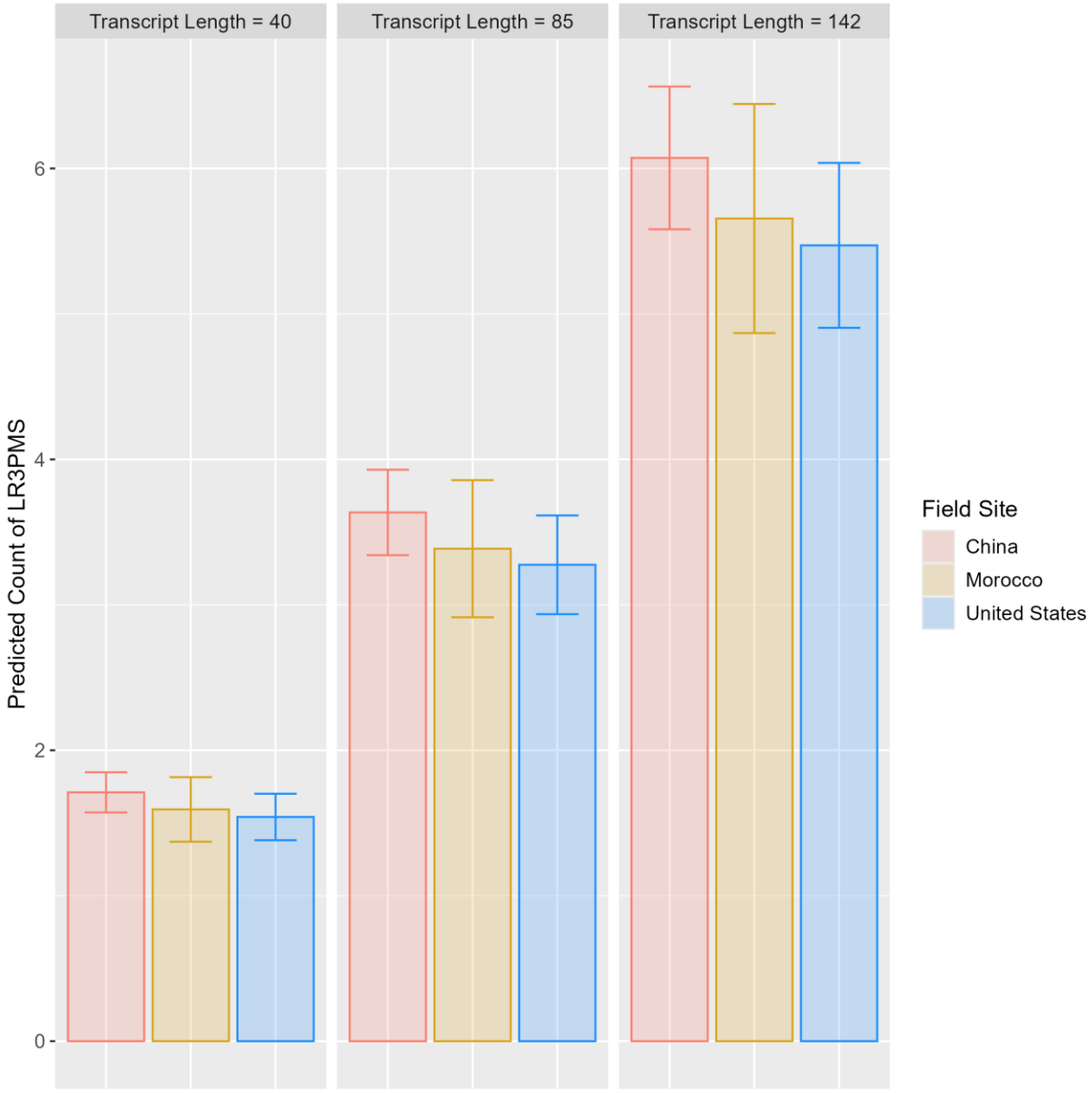
VCM 1 Predicted Estimates of Mean Counts of LR3PMS for the Lower Quartile, Median, and Upper Quartile Values of Transcript Length Across Each Level of 'Video ID'



Note. VCM 1 predicted estimates of the average count of LR3PMS organized by levels of *Video ID*. Predictions were made holding all variables except transcript length constant. Selected transcript lengths corresponded to the lower quartile, median, and upper quartile values of the variable. Error bars correspond to the 95% confidence interval.

Figure 23

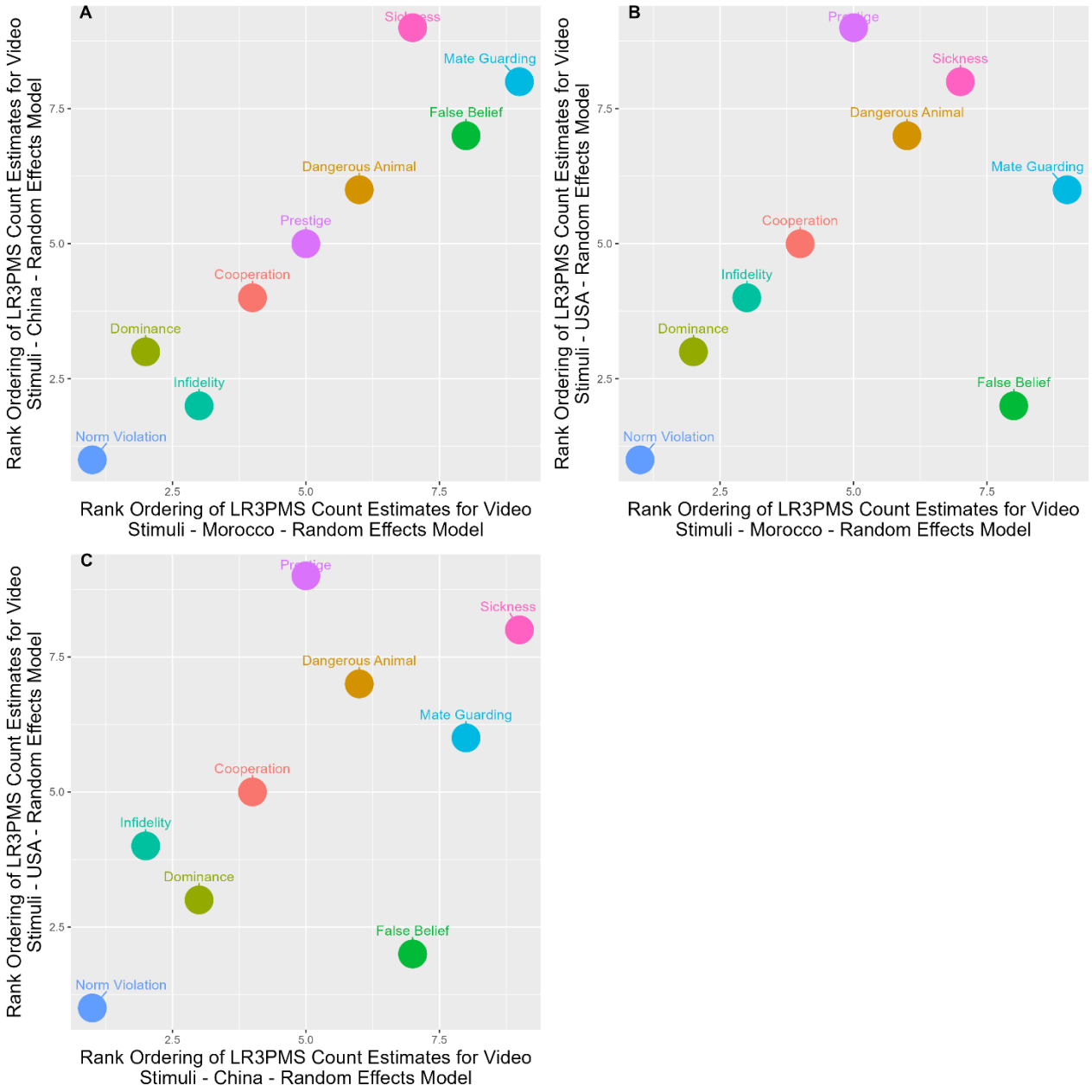
VCM 1 Predicted Estimates of Mean Counts of LR3PMS for the Lower Quartile, Median, and Upper Quartile Values of Transcript Length Across Each Level of 'Field Site'



Note. VCM 1 predicted estimates of the average count of LR3PMS organized by levels of *Field Site*. Predictions were made holding all variables except transcript length constant. Selected transcript lengths corresponded to the lower quartile, median, and upper quartile values of the variable. Error bars correspond to the 95% confidence interval.

Figure 24

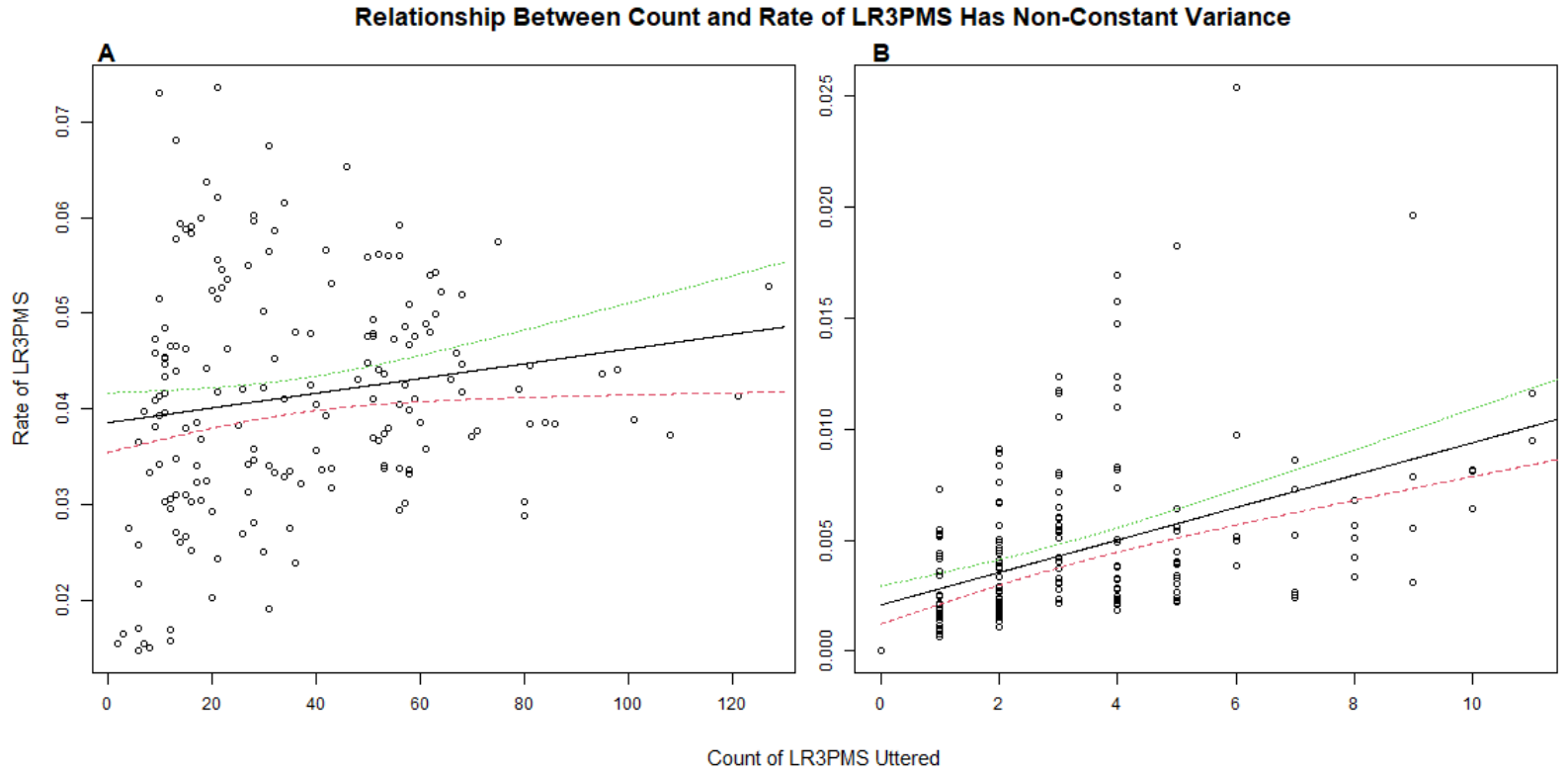
Correlation of Levels of Video ID Rank-Ordered by Predicted Count of LR3PMS in China and Morocco



Note. Points on graph corresponded to predicted counts of LR3PMS held constant at transcript lengths of 85. (A) Correlation of rank orders between China and Morocco was near-perfect and highly significant, p<.0001. (B) Despite moderate positive correlation (rho = 0.45) between rank order of the United States and Morocco, correlation was not statistically significant, p=0.2298. (C) Despite strong positive correlation (rho = 0.57) between rank order of the United States and China, correlation was not statistically significant, p=.1206.

Figure 25

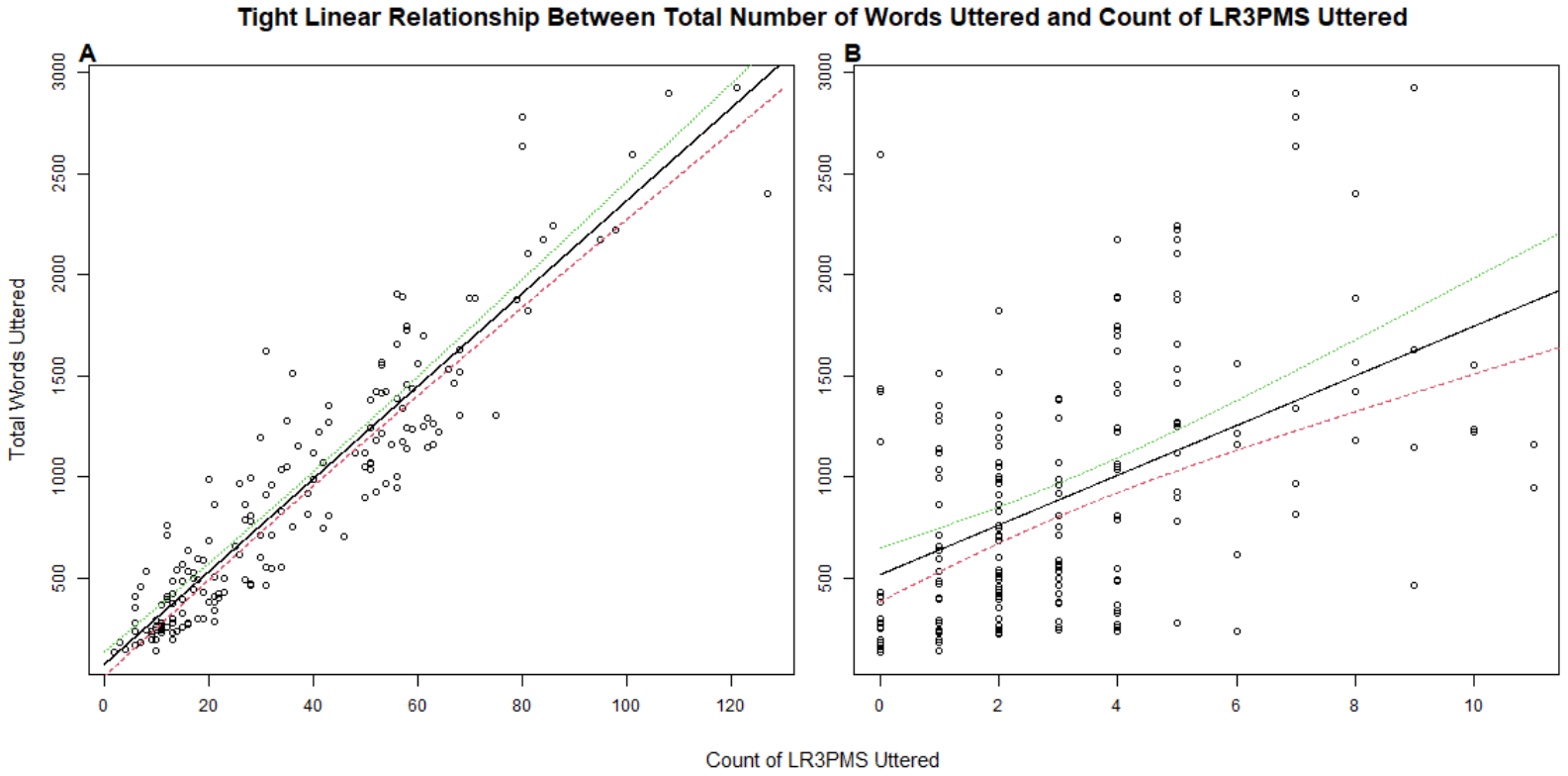
Weak Correlation Between Count and Rate of LR3PMS Uttered Has Non-Constant Variance



Note. A weak correlation exists between the *count* of LR3PMS uttered and the *rate* at which participants uttered LR3PMS. However, residuals on participant rates exhibit non-constant variation such that residuals vary more for lower counts of LR3PMS than they do for higher counts of LR3PMS. (A) LR3PMS coded using All Mental State terms. (B) LR3PMS coded using Wellman and Estes terms.

Figure 26

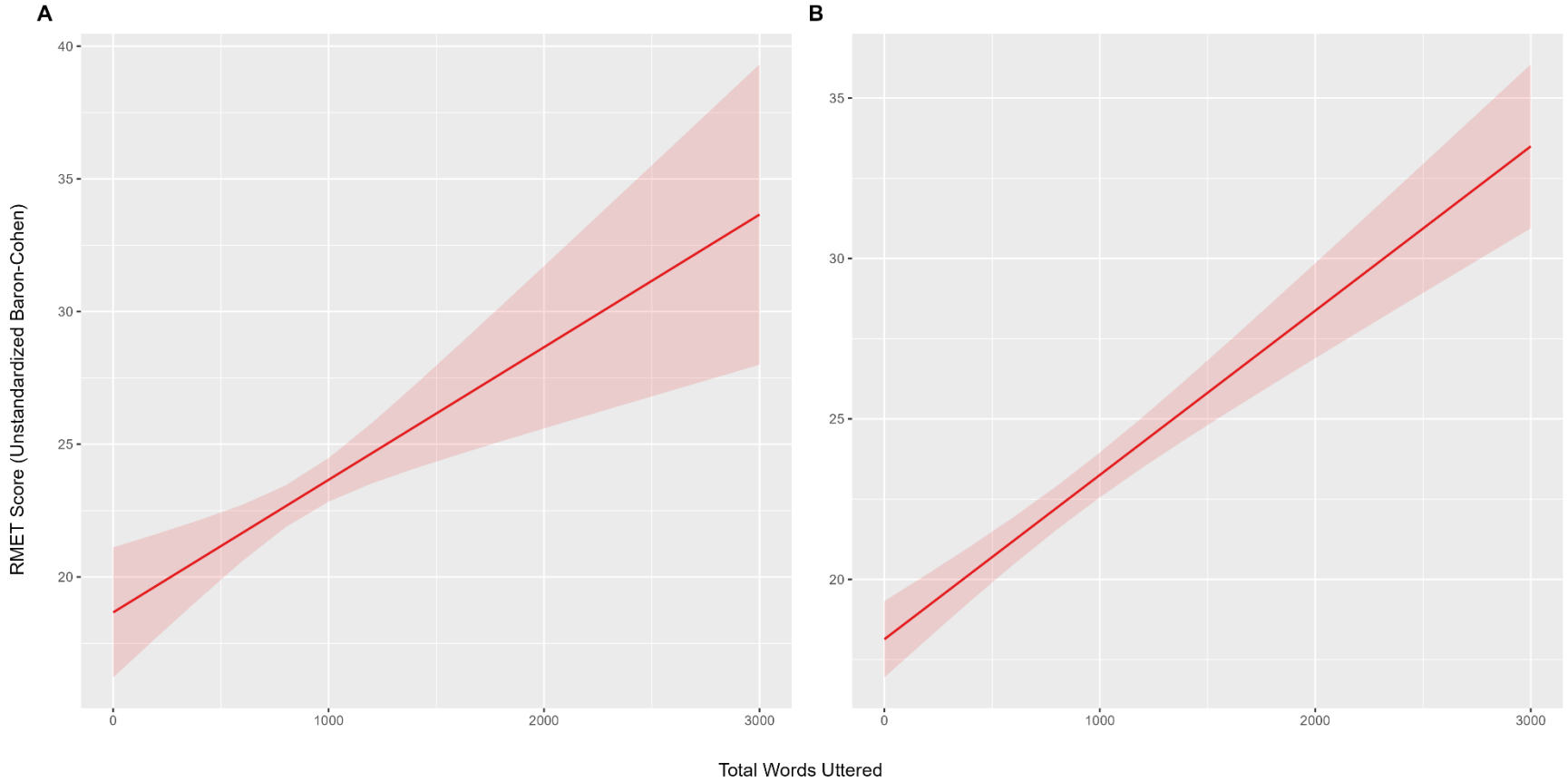
Tight Correlation Between Total Words Uttered and LR3PMS Uttered



Note. Though there are claims in the psychological literature suggesting that the count of lexical references to third-party mental states (LR3PMS) is generally dissociable from the overall quantity of speech produced, my data suggest that these two variables are highly correlated such that the more a participant spoke, the more LR3PMS they uttered. (A) LR3PMS coded using All Mental State terms. (B) LR3PMS coded using Wellman and Estes terms.

Figure 27

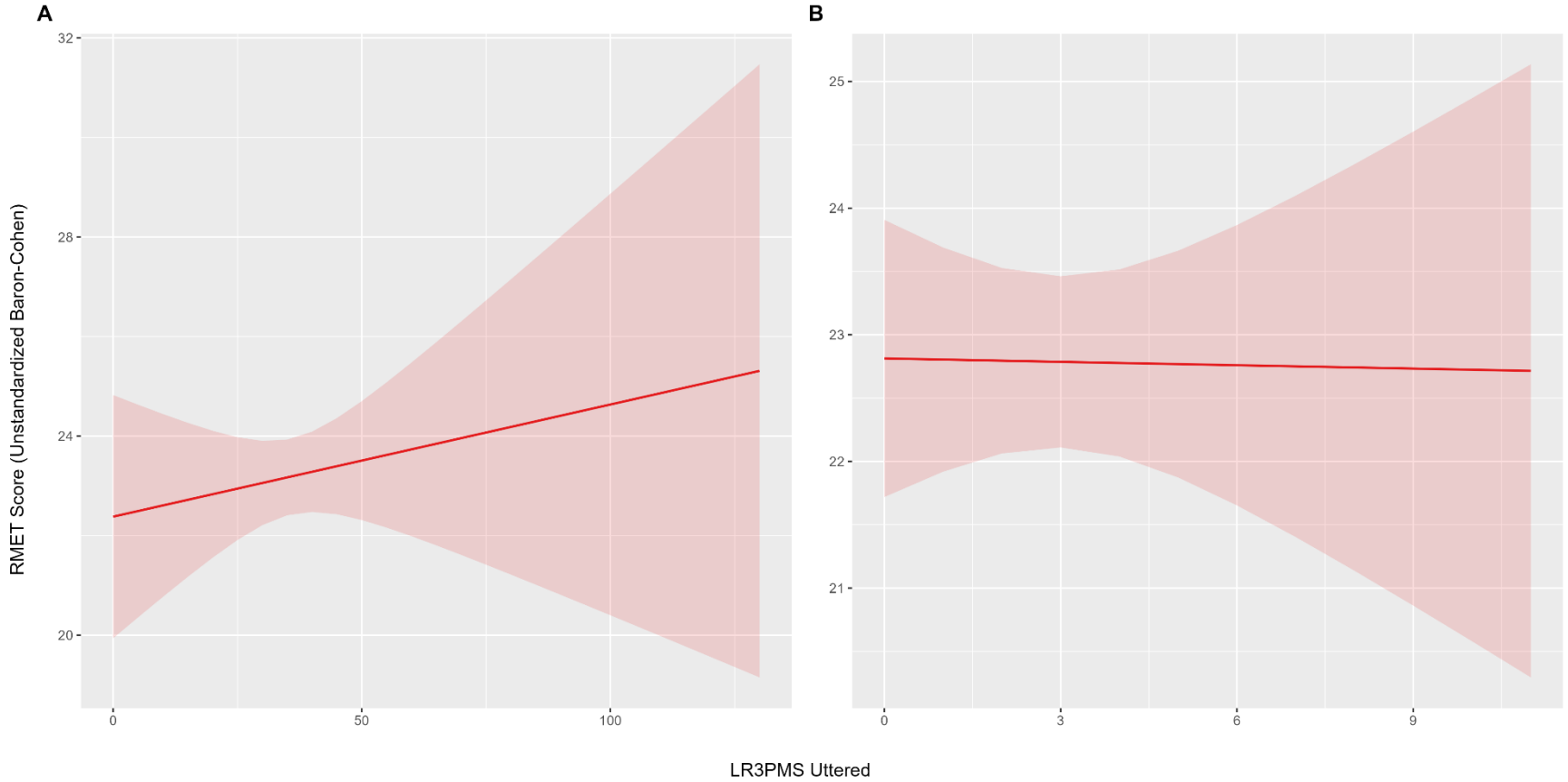
Total Words Uttered by Participants Strongly Positively Predicts Performance on RMET



Note. The predicted results of Model 8 suggest that as the counts of Total Words Uttered increase, so too do participant scores on the Reading the Mind in the Eyes Test (RMET) using the Unstandardized Baron-Cohen coding scheme. Visualization holds LR3PMS Uttered constant at the sample mean value. (A) Model 8 predictions when using LR3PMS coded using All Mental State terms. (B). Model 8 predictions when using LR3PMS coded using Wellman and Estes terms.

Figure 28

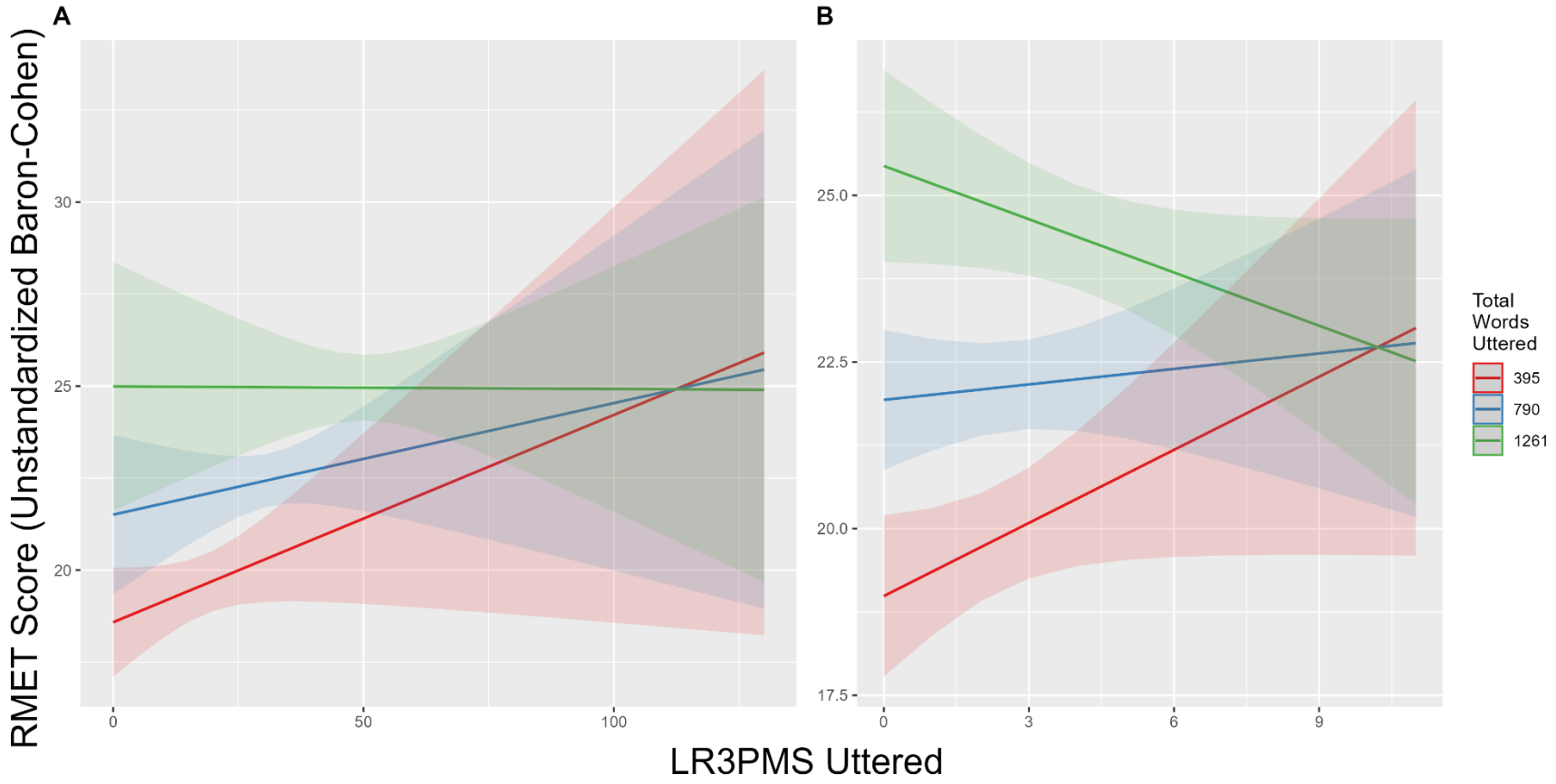
LR3PMS Uttered by Participants Predicts Performance on RMET Less Strongly Than Total Words Uttered



Note. Model 8 predicts that the participants' scores on the RMET will increase modestly as the total number of All Mental State Terms LR3PMS uttered increases. This effect is independent of, albeit weaker than that of Total Words Uttered. Visualizations hold Total Words Uttered constant at the sample mean value. (A) Model 8 predictions when using LR3PMS coded using All Mental State terms. (B). Model 8 predictions when using LR3PMS coded using Wellman and Estes terms.

Figure 29

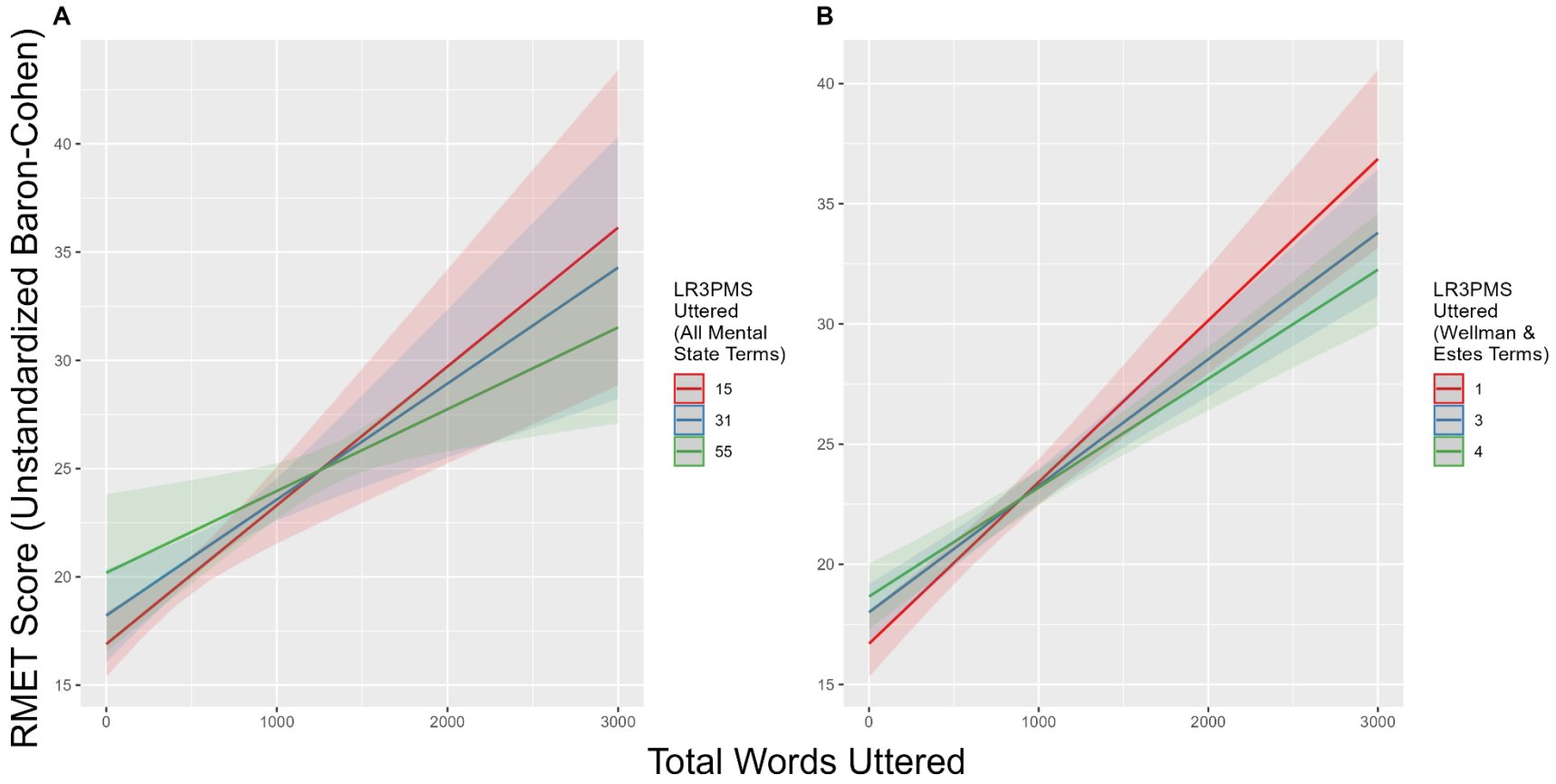
The Impact of Increased Counts of LR3PMS on RMET Scores is Attenuated as Total Words Uttered Increases



Note. (A) Predictions from Model 8 where LR3PMS were coded using All Mental State terms. (B) Predictions from Model 8 where LR3PMS were coded using Wellman and Estes terms. Values of Total Words Uttered corresponding to the lower quartile (395 words), the median (790 words), and the upper quartile (1261 words) were selected to examine the impact of increasing counts of LR3PMS uttered on RMET Score. Among the least talkative speakers, or those in the lower quartile of Total Words Uttered, as the count of LR3PMS uttered increased, performance on the RMET increased sharply (holding Total Words Uttered constant). A more modest, though still positive, effect was observed for participants who uttered the median value of Total Words Uttered. For the most talkative participants, or those in the upper quartile of Total Words Uttered, there was essentially no effect associated with a change in All Mental State term LR3PMS (A) and a negative effect with a change in Wellman and Estes term LR3PMS (B).

Figure 30

The Impact of Increased Counts of LR3PMS on RMET Scores is Attenuated as Total Words Uttered Increases



Note. *Note.* Values LR3PMS Uttered corresponding to the lower quartile (All Mental State Terms = 15; Wellman and Estes Terms = 1), the median (All Mental State Terms = 31; Wellman and Estes Terms = 3), and the upper quartile (All Mental State Terms = 55; Wellman and Estes Terms = 4) were selected to examine the impact of increasing counts of Total Words Uttered on RMET Score. Among participants who produced few LR3PMS (lower quartile), as the count of Total Words Uttered increased, performance on the RMET increased sharply (holding LR3PMS Uttered constant). A more modest, though still strongly positive, effect was observed for participants who uttered the median value of LR3PMS Uttered. For those participants who produced many LR3PMS (upper quartile), an even more modest though still fairly strongly positive effect on RMET score was observed. (A) LR3PMS coded with All Mental State terms. (B) LR3PMS coded with Wellman and Estes terms.

Appendix A

Video Questions

Cooperation Questions

- Attention Check Question: What did the first woman take out of her box?
    - Answer: Kitchen stuff
- Mental State Question 1: Why doesn't the first woman help the second woman?
    - Answer: She doesn't want to, She's too busy
- Mental State Question 2: Why does the second woman shake her head after lifting the pot?
    - Answer: She's angry, she's upset, she's in disbelief, etc.

Dangerous Animal Questions

- Attention Check Question: What animal was in the tree?
    - Answer: Snake
- Mental State Question 1: Why did the first woman stop talking and back away from the branch?
    - Answer: She noticed the snake, she was afraid of the snake, etc.
- Mental State Question 2: Why did the second woman lift the animal up with a stick after knocking it out of the tree and hitting it?
    - Answer: She was looking at it to make sure it was dead, she wanted to make sure it was dead, she thought it might still be alive, etc.

Dominance Questions

- Attention Check Question: Was the first man standing at the beginning of the video?
    - Answer: No
- Mental State Question 1: Why did the second man take the first man's wood?
    - Answer: He wanted it, he thought it was his, the other man used his ax and he was angry, he was a jerk, he wanted to intimidate him, etc.
- Mental State Question 2: Why did the second man leave a single piece of wood for the first man?
    - Answer: He was teasing him, he was insulting him, he's a jerk, etc.

False Belief Question

- Attention Check Question: What did the man leave on the tree?
    - Answer: His shirt
- Mental State Question 1: Why did the man leave after taking his shirt off?
    - Answer: He realized he forgot something, he forgot something, he had to get soap, he wasn't ready, etc.
- Mental State Question 2: Why did the man stop and look around before bathing himself?
    - Answer: He was confused, he was second-guessing himself, he thought someone stole his stuff, etc.

Infidelity Questions

- Attention Check Question: What did the second woman drop when she entered the room?
    - Answer: Her bag
- Mental State Question 1: Why did the first woman leave?
    - Answer: She wanted to give the man and the woman space, she wanted to get away, she was upset, etc.
- Mental State Question 2: Why did the man start to yell?

    - Answer: He was defending himself, he was trying to shift the blame, he felt angry, etc.

Mate Guarding Questions

- Attention Check Question: What did the second man have with him?
    - Answer: A bucket
- Mental State Question 1: Why did the first man turn the woman around and walk her away from the second man?
    - Answer: He wanted her to go, he was jealous, he wanted her to stop talking to the other man
- Mental State Question 2: Why did the woman keep talking to the second man, instead of leaving with the first man immediately?
    - Answer: She didn't realize her boyfriend was jealous, she didn't care that he was upset, etc.

Norm Violation Questions

- Attention Check Question: What gift did the women receive?
    - Answer: Flower
- Mental State Question 1: Why did the boy give the wrong gifts?
    - Answer: He was playing a prank, he wanted to mess it up, he didn't respect the ceremony, he thought it would be funny, etc.
- Mental State Question 2: Why didn't the man bow when the boy gave him his gift?
    - Answer: He was confused, he was shocked, he was in disbelief, he didn't know what to do.

Prestige Questions

- Attention Check Question: Who is a better player, the man in black or the man in pink?
    - Answer: The man in pink
- Mental State Question 1: Why does the man in black wave both his arms at the man in gray after he leaves and goes to talk to the man in pink?
    - Answer: He's trying to get his attention, he wants him to come back, he's upset that he left
- Mental State Question 2: Why does the man in black prevent the man in gray from looking at the man in pink?
    - Answer: He wants him to focus / pay attention, he wants to make sure the student learns, he notices that the student is getting distracted

Sickness Questions

- Attention Check Question: Who went to the sick person first?
    - Answer: The woman
- Mental State Question 1: Why was the sick person leaning on the tree and holding their head?
    - Answer: They felt sick, they didn't feel well, they needed support, they wanted to rest, etc.
- Mental State Question 2: Why did the man wait longer before approaching the sick person?
    - Answer: He was nervous, he didn't want to help, he was unsure about it, he wanted to get someone who knew more, etc.

Appendix B

Survey Questions

- What is your date of birth?
- What is your gender?
- What is your native language? Please list all of them if you have several native languages.
- Do you identify as a speaker of a particular dialect or variety of this language (or languages)?
- What is this dialect called, or where is it mostly spoken?
- What is your nationality?
- In which country do you live?
- In which country were you born?
- Are you a student?
- What is your profession or occupation?
- What is your education level?
- How much philosophy have you done during your education?
- To the best of your knowledge, what is the education level of your parent who has had the most education?
- What is your current religious affiliation(s)?
- How important is religion in your life?
- What is your general political attitude?
- How many siblings did you grow up with?
- How many close friends would you say you have?

Appendix C

Supplement to Chapter 2

Methods

Procedure

Data collection. The purpose of these pseudorandomized conditions allowed me to minimize potential order effects of video presentation on participant descriptions, conditional on some of the other study goals; specifically, to pilot-test sets of targeted questions about each video after participants completed their descriptions. These questions were meant to directly probe participants' attributions of mental states to the characters depicted in the video stimuli, thus introducing the risk of cuing participants to the study's purpose. To pilot these questions while ensuring a subset of participants' descriptions were immune to this risk, videos in each of the pseudorandomized conditions were broken into two blocks and questions followed only those narrative descriptions in the second block. Finally, the false belief video was fixed at the end of one of the two blocks across pseudorandomized conditions, with the false belief video following the first 4 videos comprising block 1 in conditions 1 – 4 and following the second 4 videos comprising block 2 in conditions 5 – 8. This left narrative descriptions of the false belief video from half of the participants recruited at each field site to be unaffected by possible cuing.

Data processing. Following data collection, participant responses were processed into a form amenable to coding for the occurrence of LR3PMS. Across both field sites and distinct implementations of the study software, three data formats were generated – ODK Collect forms, virtual interview forms with participant audio clips, and virtual interview forms without participant audio clips. Each format had distinct processing demands in order to produce transcripts research assistants fluent in each of the target languages could correct and standardize into a long-form data format amenable to coding for LR3PMS.

ODK Collect outputs. Raw data outputs from ODK Collect comprise a single, wide-format spreadsheet wherein each row corresponds to a single participant and each column corresponds to distinct “question” items. Note that “question” items do not entail questions per se; rather, “question” items are the minimal units of survey construction in the ODK Collect environment. Elements like experimenter scripts are “question” items insofar as they occur in the raw data output as distinct columns, albeit absent responses in their corresponding rows. While such cells could be deleted by hand to transform the data into long-format documents, the raw ODK collect data presented a number of other constraints disfavoring this course of action.

As indicated above in the *Data collection* section, ODK collect does not permit randomization of question order. Thus, each survey included all eight pseudorandomized conditions, though the conditions to which the participant was not assigned were hidden during survey presentation. Each row in the data thus contained wide swaths of empty cells corresponding to those conditions. Since question order differed across conditions, converting the data into long-form would remain difficult, even if not for these varying sets of blank cells.

A final concern necessitating the data be processed by Python script is the organization and management of the four data types collected in the survey. These data were collected as part of a broader research program to analyze the qualities, causes, and consequences of LR3PMS. As such, they consist of more than just narrative descriptions of video stimuli and include participant demographics, links to WAV files of participants’ descriptions, experimenter evaluations of the correctness and mind-mindedness of participants’ responses to question sets following videos in the second block of each pseudorandomization condition, and finally participant responses to the RMET. I wanted to produce separate documents for each of these data types, though they were interlaced in the raw ODK outputs. Thus, Python scripts were written to parse the raw data accordingly.

Four distinct scripts were written to process the ODK Collect data. The first of these scripts served to crawl through the cells of each row looking for hyperlinks to the WAV files stores on the remote ONA server. Upon identifying a cell containing a hyperlink, the script opened the link and downloaded the file into a folder specified by the user. Folders were created for each field site and files were saved correspondingly. Additionally, the script extracted descriptive information about the file from the spreadsheet subject ID, video ID, the order in which the file was recorded, and the pseudorandomization condition and renamed the audio file accordingly. Once all WAV files for which there were hyperlinks in the raw data had been downloaded and renamed, preliminary transcripts for each WAV file were generated using a script written to interface with the Google Speech-to-Text API. As with the previous script, users specified where to save the transcripts and titles were assigned so as to match the corresponding audio. This script also permitted the user to specify the language of the audio files (although the options were limited to those for which Google's Speech-to-Text API provides support). The text within a transcript was written in the standard orthography of the corresponding language.

Next, three scripts were written to produce spreadsheets corresponding to question data, word data, and Reading the Mind in the Eyes Test data. The first of these scripts was written to produce a long-form datasheet containing experimenter evaluations of participant questions responses, wherein each row corresponded to a single evaluation (either correct / incorrect or contains mental states / does not contain mental states) coupled with pertinent descriptors of the data point and participant demographics. Attention check questions were evaluated only as correct / incorrect, whereas mental-state questions were each evaluated according to their correctness and for the presence or absence of mental states. The second of these scripts was written to extract participants' demographic data and create a long-form data sheet of participant responses to each of the 36 items (and 1 trial item) in the Reading the Mind

in the Eyes Test. Each row in this spreadsheet corresponded to a single item response coupled with pertinent descriptors of the data point and participant demographics.

Outputs of custom software for virtual interviews (functioned correctly). Given the timeline of the data collection and how it corresponded with that of the global COVID-19 pandemic, the post-processing data structures described in the preceding section were already known, and thus the software was written in such a way as to generate WAV files, transcripts, and data sheets of question and Reading the Mind in the Eyes data as part of its function. As such, experimenters using this software needed only to specify the language in which the study was to be run and the location where they would like to save its outputs and run the study, at which point it would generate a structured set of folders housed in the location specified at the outset. This is all to say that in cases where the software worked as intended, no additional processing unique to this approach was required for progressing to the next stage of processing.

Outputs of custom software for virtual interviews (functioned incorrectly). When the custom software did not function as intended, it failed in a characteristic manner. Namely, all datasheets were generated without error. However, the audio clips that the program generated were silent and the corresponding transcripts were blank. These cases are attributable to failure on the part of the experimenter to modify their computer's audio settings to accommodate the mechanism by which participant speech audio was captured and written to WAV files. Effectively, the software was designed to loopback speaker audio – that is, to treat audio output driven through a computer's speakers as if it were audio input fed into a microphone. This procedure requires the modification of a few audio settings, including enabling of a digital loopback audio "device" and resetting the device's default audio input and output. When the software was given the command to begin recording, it would search the computer's audio inputs and select the stream of audio handled by the loopback device. With this audio stream

selected, it would then create a WAV file into which the speaker audio would be written, stopping only when indicated by the user. Due to variability in default audio settings across individual computers, experimenters met with the study lead to troubleshoot and ensure their access to at least one computer capable of capturing speaker output. Experimenters were told not to use any other computers to run the study unless they were subject to such troubleshooting too. Despite these precautions, a subset of these data was returned wherein the WAV files had not captured participant audio and, correspondingly, the transcripts were blank. Crucially, the experimental procedure also dictated that audio and video of the interview be captured using the videoconferencing software of the experimenter's choice. In these instances, audio files were processed by either of the following methods. Audio files of the full-length interview were processed into individual clips using Audacity, a free and open-source audio recording and editing software. Clips were named according to the same scheme described above, saved into individual folders, and subsequently transcribed using a custom Python script that interfaced with the Google Speech-to-Text API where there existed language support. In other cases, full-length audio files were fed into Adobe Premiere in order to generate transcripts. Research assistants fluent in the target language edited the transcript in order to create sections that corresponded to participant responses, and correct errors, after which point the full-length transcript was exported as a text file and edited into its separate text files. These were then named using the earlier-described naming scheme.

Transcript editing. Once data across all formats (ODK Collect outputs, outputs from custom software written for virtual interviews that operated as intended, and outputs from custom software written for virtual interviews that did not operate as intended) had undergone any necessary, format-specific processing, data from each collaborating field site were now structured as a library of transcripts stored in simple .TXT files. Prior to final processing and coding, transcripts were reviewed and corrected due to inconsistencies in the quality of machine

transcription across languages. These differences are attributable both to limitations in the size of the training corpus on which such proprietary machine-learning algorithms (such that some languages, like English, have much larger training data sets than languages like Moroccan Arabic, and thus produce more accurate transcripts of novel audio), as well as differences across collaborating field sites in audio quality. Therefore, no fewer than two bilingual research assistants fluent in English as a second language and one of the four target languages as their first, and literate in the standard orthography of their first language, were recruited to review and correct the transcripts.

Research assistants were provided with all transcripts derived from interviews conducted in their first language by experimenters at one of four collaborating field sites, as well as all corresponding WAV files from which the transcripts were generated. For a given transcript, research assistants were instructed to open both the text file and the corresponding WAV file. Research assistants would then read the text file of the transcript as the WAV file played, pausing as necessary to correct typos, errors of commission (words that were not uttered by the participant but included in the transcript), and errors of omission (words that were uttered by the participant but not included in the transcript). More specifically, research assistants made these edits to the transcripts according to the following criteria:

- Ensure that spelling is correct.
- Ensure that all completely uttered words, including repetitions (the, the man) and filler words (uh, um, hmm), are included.
- Where self-interruptions or incomplete terms occur (i.e., "I s-, I saw"), include only those words that are complete (I, I saw).
- Non-linguistic utterances including laughter, sighs, coughs, and so forth were not transcribed .

- Research assistants did not need to amend or include punctuation in the transcripts. But, in cases where punctuation was included, they were asked to delete it due to additional downstream data processing requirements.
- For languages in which the standard orthography did not include space characters between words, research assistants were asked to add spaces between words, where words were defined as the minimal linguistic units necessary to capture complete concepts.

Creating documents for coding data. After all transcripts had been reviewed and edited by research assistants, the final stage of data processing before coding was performed. In this stage, TXT files of the transcripts were processed into two separate spreadsheets. The first of these spreadsheets contained all of the words, in sequential order, in all of the transcripts for a given language as long-form data such that each row corresponded to a single word coupled with indicators of subject ID, video ID, pseudorandomization condition, and order of presentation in addition to demographic data from the corresponding participant. This spreadsheet was generated using a custom Python script that appended rows to the spreadsheet sequentially, such that, for a given row, the proceeding row corresponded to the next word in the transcript and the preceding row corresponded to the previous word from the same transcript. When all words in a transcript had been processed by the script in this manner, the script proceeded to the next transcript, updating the indicators and demographic data as necessary before appending the row to the spreadsheet. At these junctures, the preceding row constituted the last word of the previous transcript. Although the data at this stage had undergone significant processing, this spreadsheet was referred to as the “Raw Data” spreadsheet by the lead author and research assistants for ease of communication.

The second spreadsheet was then generated using another Python script that surveyed each row in the Raw Data spreadsheet to identify all unique word types and counts of their

tokens across rows. The resulting document consisted of three columns. The first contained all word types in the Raw Data spreadsheet, organized alphabetically. The second contained token count, or word frequency. The third column was left intentionally blank, as research assistants would ultimately code this document for the occurrence of those unique elements that could be glossed as referring to third-party mental states. This document was referred to as the "Dictionary" spreadsheet by the lead author and research assistants due to its structure roughly mirroring that of a frequency dictionary.

Supplement to Chapter 3

Results

Variance Component Model 2 (VCM 2)

As the full model against which the remaining two would be compared, VCM 2 was fit in order to examine the role of *Video ID,* the role of *Field Site,* and the role of their interaction simultaneously. VCM 2 fit was evaluated using the AIC, or Akaike Information Criterion (AIC = 1972.5) and the BIC, or Bayesian Information Criterion (BIC = 199.3). The log-likelihood of the model was also reported (log-likelihood = -981.2). Variance estimates and standard deviations for each predictor in the model were assessed to determine the variability the random effects captured, the results of which can be found in **Table S1**. Predictors are referred to by their variable names in plain English rather than using the syntax of the model to which they corresponded. Variances correspond to the spread in the intercepts among between the levels of each variable. Therefore, the variance of the random effect may be understood as the degree to which the intercepts corresponding to each level within a given variable vary. The lower this number, the less variation present across the levels of the variable.

With this interpretation in mind, the variables for which variation was greatest between levels were *Video ID* (var = 0.56473, sd = 0.715) and interaction between *Video ID* and *Field Site* (var = 0.4947, sd = 0.7034), respectively. Crucially, the variance estimates for the remaining variables exhibited two features of note. First, the variance estimate for *Participant ID* (var = 0.12428, sd = 0.3525) was substantially greater than that of *Field Site* alone (var = 0.02472, sd = 0.1572). The second is that both of these values were substantially lower than either of the estimates for *Video ID* or for the interaction between *Video ID* and *Field Site.* Intraclass Correlation Coefficients were calculated for each random effect to determine the proportion of the total variance explained by each. 10.28% of the total variance explained by the model was attributable to *Participant ID*, 46.73% was attributable to the interaction between *Field Site* and

Video ID, 40.94% was attributable to *Video ID*, and only 2.05% was attributable to *Field Site*. **Figures S1, S2,** and **S3** illustrate the conditional modes of the random intercepts estimates with 95% confidence intervals for each of the variables included in the model, untransformed. The conditional modes correspond to the deviation of a specific group's intercept from the overall average intercept, conditional on the data. In essence, then, the values presented in **Figures S1 – S3** represent simultaneously the extent to which each level of the variable differs from the overall average intercept and the extent to which these levels differ from each other.

As can be seen in **Figure S1**, the 95% confidence intervals for all three field sites overlap substantially with each other and include zero. Collectively, these results indicate that the conditional modal estimates for each level of *Field Site* differ neither from each other nor from the intercept estimated for the overall average. In **Figure S2**, the 95% confidence intervals for each of the nine video stimuli indicate that the only video stimulus for which the random intercept is reliably different from that over the overall average intercept is the False Belief video, though the 95% confidence interval for the Mate Guarding video overlaps with zero only very slightly. The confidence intervals on the conditional modal intercept estimate for the Mate Guarding video overlaps with those of every other video, while the confidence intervals on the conditional modal intercept estimate for the False Belief video is reliably different than those of the Dangerous Animal and Cooperation videos. Taken together, these results suggest that at least one of the video stimuli, though possibly two tended to elicit a greater number of LR3PMS than average. Furthermore, these results support the conclusion that the count of LR3PMS in transcripts describing the False Belief video stimulus was reliably higher than in transcripts describing the Dangerous Animal and Cooperation video stimuli.

In **Figure S3A,** it can be seen that none of the conditional modal estimates of the random intercepts for *Video ID* reliably differed from zero or from each other among participants recruited from China. The same can generally be said for participants recruited from the United

States, with the exception of the conditional modal estimate of the intercept for the Cooperation video stimulus which was found to be reliably below average. In contrast, three of the conditional modal estimates of the random intercepts for *Video ID* differed reliably or nearly reliably from zero among participants recruited from Morocco. Among these participants, estimates were higher than the overall average intercept estimate for the False Belief video stimulus and the Prestige video stimulus, while nearly reliably lower for the Sickness video stimulus. **Figure S3B** presents the same data as **Figure S3A** grouped by *Video ID* on the y axis and helps to illustrate that across all 9 video stimuli, the conditional modal estimates of the random intercepts for each country do not reliably differ from each other across any of the video stimuli.

Supplement to Chapter 4

Results

Variance Component Model 2 (VCM 2)

As the full model against which the remaining two would be compared, VCM 2 was fit in order to examine the role of *Participant ID,* the role of *Video ID,* the role of *Field Site,* and the role of their interaction simultaneously. VCM 2 fit was evaluated using the AIC, or Akaike Information Criterion (AIC = 6143.2) and the BIC, or Bayesian Information Criterion (BIC = 6170.0). The log-likelihood of the model was also reported (log-likelihood = -3066.6). Variance estimates and standard deviations for each predictor in the model were assessed to determine the variability captured by the random effects, the results of which can be found in **Table S2**. Predictors are referred to by their variable names in plain English rather than using the syntax of the model to which they corresponded. Variances correspond to the spread in the intercepts across the levels of each variable. Therefore, the variance of the random effect may be understood as the degree to which the intercepts corresponding to each level within a given variable vary. The lower this number, the less variation present across the levels of the variable.

With this interpretation in mind, the variables for which variation was greatest between levels were *Video ID* (var = 0.0915, sd = 0.3025) and the interaction between *Video ID* and *Field Site* (var = 0.05245, sd = 0.2290), respectively. Crucially, the variance estimates for the remaining variables exhibited three features of note. First, the variance estimate for *Participant ID* (var = 0.03701, sd = 0.1924) was substantially greater than that of *Field Site* alone (var < 0.0001, sd < 0.0001). Second, both of these values were substantially lower than either of the estimates for *Video ID* or for the interaction between *Video ID* and *Field Site.* Lastly, the variance estimate for *Field Site* alone was effectively zero. This state of affairs is the consequence of singular model fit, or cases wherein some dimensions of the variance-covariance matrix have been estimated as exactly or very nearly zero. While models with

singular fit are statistically well-defined, as it is theoretically sensible for the true maximum likelihood estimate to correspond to a singular fit, such fits may correspond to overfitted models with poor power. This problem is one to which I will return later in the discussion to adjudicate whether this indicates poor model fit and whether how it ought to be understood in light of model-building considerations. Intraclass Correlation Coefficients were calculated for each random effect to determine the proportion of the total variance explained by each. 20.45% of the total variance explained by the model was attributable to *Participant ID*, 28.98% was attributable to the interaction between *Field Site* and *Video ID*, 50.56% was attributable to *Video ID*, and effectively none of the variance was attributable to *Field Site*. **Figures S4, S5, S7,** and **S8** illustrate the conditional modes of the random intercepts estimates with 95% confidence intervals for each of the variables included in the model, untransformed. The conditional modes correspond to the deviation of a specific group's intercept from the overall average intercept, conditional on the data. In essence, then, the values presented in **Figures S5 – S8** represent simultaneously the extent to which each level of the variable differs from the overall average intercept and the extent to which these levels differ from each other.

As can be seen in **Figure S4**, the 95% confidence intervals for all three field sites overlap entirely with each other and are centered on zero. Collectively, these results indicate that the conditional modal estimates for each level of *Field Site* differ neither from each other nor from the intercept estimated for the overall average. In **Figure S5**, the 95% confidence intervals for each of the nine video stimuli indicate that the only random intercepts reliably different from that of the overall average intercept are those for the Sickness and Norm Violation videos, though the 95% confidence interval for the Prestige and Mate Guarding videos overlap with zero only very slightly. The confidence intervals on the conditional modal intercept estimate for the Mate Guarding, Prestige, and Sickness videos overlap with those of every other video, while the confidence intervals on the conditional modal intercept for the Norm Violation video is

reliably different from every video other than Infidelity and Dominance. Taken together, these results suggest that at least one of the video stimuli tended to elicit fewer LR3PMS than the overall average and at least one of the video stimuli tended to elicit more LR3PMS than the overall average.

In **Figure S7A,** it can be seen that none of the conditional modal estimates of the random intercepts for *Video ID* reliably differed from zero or from each other among participants recruited from China. The same can generally be said for participants recruited from the United States, with the exception of the conditional modal estimate of the intercept for the False Belief video stimulus which was found to be reliably below average and the Prestige video stimulus which was found to be reliably above average. In contrast, four of the conditional modal estimates of the random intercepts for *Video ID* differed reliably or nearly reliably from zero among participants recruited from Morocco. Among these participants, estimates were higher than the overall average intercept estimate for the False Belief video stimulus and nearly reliably higher for Mate Guarding video stimulus, while estimates were lower than the overall average intercept estimate for both the Norm Violation and Dominance video stimuli. **Figure S7B** presents the same data as **Figure S7A** grouped by *Video ID* on the y axis and helps to illustrate that across all 9 video stimuli, the conditional modal estimates of the random intercepts for each country do not reliably differ from each other across any of the video stimuli with the exception of the Dominance video stimulus.

Supplement to Chapter 5

Results

Descriptive Statistics and Assumption Checking

As found in Chapter 2, both the total number of LR3PMS and the total number of words uttered by participants are Poisson-distributed statistics. However, in Chapter 2, the former of these two statistics was treated as a dependent variable and the latter was treated as a log offset term. In the present analysis, I treated them both as predictor variables in my model. I examined the distributions of all non-categorical predictors to ascertain their shape and pick an appropriate modeling strategy. Plots of these distributions can be found in **Figures S8 – S17**.

As can be seen, the distributions in **Figures S8, S9,** and **S10** are non-normal. This is perhaps unsurprising given the process according to which word count data is generated. As count data, there is a lower limit on values each of these variables can take such that no participant can produce fewer than 0 words. Similarly, most participants tend to produce roughly the same number of words, and it is only rarely that a participant speaks extensively. Collectively, these processes generate data that are more or less Poisson distributed. Interestingly, the distribution of counts of LR3PMS appears to be bimodally distributed, with peaks occurring at counts of approximately 15 and 50 when using the All Mental State Terms coding scheme and with peaks occurring at counts of 3 and 9 when using the Wellman and Estes Terms coding scheme, though addressing this pattern is outside the scope of the current project. Having characterized these first two parameters, it is notable that the per-participant rate at which LR3PMS are generated when using the All Mental State Terms coding scheme appears to be approximately normally distributed (**Figure S11**) In contrast, the per-participant rate at which LR3PMS are generated when using the Wellman and Estes Terms coding scheme appears to be approximately Poisson-distributed (**Figure S12**). All four versions of the RMET

scores exhibited similarly approximately normally distributed data (**Figures S13 – S16**). These patterns of data provided a useful foundation upon which to build my analytic strategy.

Model Comparison and Selection

Model selection and analysis for Standardized Baron-Cohen RMET Scores

LR3PMS Uttered (Wellman and Estes Terms). In all 15 models, simple linear regression was used to determine if the predictor variables significantly predicted Standardized Baron-Cohen RMET scores. After running all 15 models, Model 8 was found to have the lowest AIC score (AIC: 433.653) and predicted Standardized Baron-Cohen RMET scores from Total Words Uttered, LR3PMS Uttered, and the interaction between Total Words Uttered and LR3PMS Uttered. Model 8 was found to be statistically significant in its overall fit ($R^2 = 0.3477$, $F(3, 173) = 30.74$, $p < .0001$). It was found that Total Words Uttered ($\beta = 0.001488$, $p < .0001$), LR3PMS Uttered ($\beta = 0.1455$, $p = .004$), and their interaction ($\beta = -.000157$, $p < .0001$) all significantly predicted Standardized Baron-Cohen RMET scores. Thus, for a single-word increase in the count of Total Words uttered, the best-fit model predicts RMET scores to increase by approximately 0.0015 standard deviations. As standardized RMET scores are not intrinsically interpretable on their own, it is perhaps more meaningful to frame this result in terms of the change in Total Words Uttered expected to correspond to a score one standard deviation above the mean. Characterized like this, RMET scores are expected to increase by one standard deviation above the population mean for an increase of 673 words in the count of Total Words Uttered. For a single-word increase in the count of LR3PMS uttered, the best fit model predicts RMET scores to increase by approximately 0.1455 standard deviations. As before, this means that RMET scores are expected to increase by one standard deviation above the population mean for an increase of seven words in the count of LR3PMS uttered. Finally, the interaction term can be understood as follows. For a single-word increase in the count of Total Words Uttered, each single-word increase in the count of LR3PMS uttered is expected to

decrease RMET scores by 0.000157 standard deviations. **Figure S19A** illustrates the predicted Standardized Baron-Cohen RMET scores with 95% confidence intervals across the range of values for Total Words Uttered. **Figure S20A** illustrates the predicted Standardized Baron-Cohen RMET scores with 95% confidence intervals across the range of values for LR3PMS. **Figure S21A** illustrates the effect of LR3PMS Uttered on Standardized Baron-Cohen RMET scores with 95% intervals for participants with a Total Words Uttered count of 395 (the value corresponding to the lower quartile of words uttered), 790 (the median number of words uttered), and 1261 words (the value corresponding to the upper quartile of words uttered).. **Figure S22A** is essentially the same as **Figure S21A**, though the x-axis now corresponds to the count of Total Words Uttered. **Figure S22A** illustrates the effect of Total Words Uttered on Standardized Baron-Cohen RMET scores with 95% intervals for participants with an LR3PMS count of 1 (the value corresponding to the lower quartile of LR3PMS uttered), 3 (the median number of LR3PMS uttered), and 4 words (the value corresponding to the upper quartile of LR3PMS uttered). **Figure S23** contains visualizations of the fit statistics of Model 8.

LR3PMS Uttered (All Mental State Terms). In all 15 models, simple linear regression was used to determine if the predictor variables significantly predicted Standardized Baron-Cohen RMET scores. After running all 15 models, Model 8 was found to have the lowest AIC score (AIC: 429.746) and predicted Standardized Baron-Cohen RMET scores from Total Words Uttered, LR3PMS Uttered, and the interaction between Total Words Uttered and LR3PMS Uttered. Model 8 was found to be statistically significant in its overall fit ($R^2 = 0.3509$, $F(3, 173) = 32.71$, $p < .0001$). It was found that Total Words Uttered ($\beta = 0.001572$, $p < .0001$), LR3PMS Uttered ($\beta = 0.01559$, $p = .0332$), and their interaction ($\beta = -.0000142$, $p < .0001$) all significantly predicted Standardized Baron-Cohen RMET scores. Thus, for a single-word increase in the count of Total Words uttered, the best-fit model predicts RMET scores to increase by approximately 0.0016 standard deviations. As standardized RMET scores are not intrinsically

interpretable on their own, it is perhaps more meaningful to frame this result in terms of the change in Total Words Uttered expected to correspond to a score one standard deviation above the mean. Characterized like this, RMET scores are expected to increase by one standard deviation above the population mean for an increase of 637 words in the count of Total Words Uttered. For a single-word increase in the count of LR3PMS uttered, the best fit model predicts RMET scores to increase by approximately 0.01559 standard deviations. As before, this means that RMET scores are expected to increase by one standard deviation above the population mean for an increase of 65 words in the count of LR3PMS uttered. Finally, the interaction term can be understood as follows. For a single-word increase in the count of Total Words Uttered, each single-word increase in the count of LR3PMS uttered is expected to *decrease* RMET scores by 0.000014 standard deviations. **Figure S19B** illustrates the predicted Standardized Baron-Cohen RMET scores with 95% confidence intervals across the range of values for Total Words Uttered. **Figure S20B** illustrates the predicted Standardized Baron-Cohen RMET scores with 95% confidence intervals across the range of values for LR3PMS. **Figure S21B** illustrates the effect of LR3PMS Uttered on Standardized Baron-Cohen RMET scores with 95% intervals for participants with a Total Words Uttered count of 395 (the value corresponding to the lower quartile of words uttered), 790 (the median number of words uttered), and 1261 words (the value corresponding to the upper quartile of words uttered). **Figure S22B** is essentially the same as **Figure S21B**, though the x-axis now corresponds to the count of Total Words Uttered. **Figure S22B** illustrates the effect of Total Words Uttered on Standardized Baron-Cohen RMET scores with 95% intervals for participants with an LR3PMS count of 1 (the value corresponding to the lower quartile of LR3PMS uttered), 3 (the median number of LR3PMS uttered), and 4 words (the value corresponding to the upper quartile of LR3PMS uttered). **Figure S24** contains visualizations of the fit statistics of Model 8.

Model selection and analysis for Unstandardized Culturally Variable RMET Scores

LR3PMS Uttered (Wellman and Estes Terms). In all 15 models, simple linear regression was used to determine if the predictor variables significantly predicted Unstandardized Culturally Variable RMET scores. After running all 15 models, Model 8 was found to have the lowest AIC score (AIC: 1000.735) and predicted Unstandardized Culturally Variable RMET scores from Total Words Uttered, LR3PMS Uttered, and the interaction between Total Words Uttered and LR3PMS Uttered. Model 8 was found to be statistically significant in its overall fit (R^2 = 0.3088, $F(3, 173) = 27.21$, $p < .0001$). It was found that Total Words Uttered (β = 0.006585, p < .0001), LR3PMS Uttered (β = 0.5515, p = .0271), and their interaction (β = -.0006064, p = .00175) all significantly predicted Unstandardized Culturally Variable RMET scores. Thus, for a single-word increase in the count of Total Words uttered, the best-fit model predicts Unstandardized Culturally Variable RMET performance to increase by approximately 0.007 points. As the RMET is scored on an integer scale ranging from 0 to 36, it is perhaps more meaningful to frame this result in terms of the change in Total Words Uttered expected to correspond to a single-point increase in Unstandardized Culturally Variable RMET score. Characterized like this, a single-point increase in Unstandardized Culturally Variable RMET score is expected for each increase of 152 words in Total Words Uttered. For a single-word increase in the count of LR3PMS uttered, the best fit model predicts Unstandardized Culturally Variable RMET scores to increase by approximately 0.55 points. As before, this means that for a 2 word increase in the count of LR3PMS uttered, Unstandardized Culturally Variable RMET score is expected to increase by approximately 1 point. Finally, the interaction term can be understood as follows. For a single-word increase in the count of Total Words Uttered, each single-word increase in the count of LR3PMS uttered is expected to *decrease* Unstandardized Culturally Variable RMET scores by 0.00061 points. **Figure S25A** illustrates the predicted Unstandardized Culturally Variable RMET scores with 95% confidence intervals across the range of values for Total Words Uttered. **Figure S26A** illustrates the predicted Unstandardized

Culturally Variable RMET scores with 95% confidence intervals across the range of values for LR3PMS. **Figure S27A** illustrates the effect of LR3PMS Uttered on Unstandardized Culturally Variable RMET scores with 95% intervals for participants with a Total Words Uttered count of 395 (the value corresponding to the lower quartile of words uttered), 790 (the median number of words uttered), and 1261 words (the value corresponding to the upper quartile of words uttered). **Figure S28A** is essentially the same as **Figure S27A**, though the x-axis now corresponds to the count of Total Words Uttered. **Figure S28A** illustrates the effect of Total Words Uttered on Unstandardized Culturally Variable RMET scores with 95% intervals for participants with an LR3PMS count of 1 (the value corresponding to the lower quartile of LR3PMS uttered), 3 (the median number of LR3PMS uttered), and 4 words (the value corresponding to the upper quartile of LR3PMS uttered). **Figure S29** contains visualizations of the fit statistics of Model 8.

LR3PMS Uttered (All Mental State Terms). In all 15 models, simple linear regression was used to determine if the predictor variables significantly predicted Unstandardized Culturally Variable RMET scores. After running all 15 models, Model 8 was found to have the lowest AIC score (AIC: 994.2647) and predicted Unstandardized Culturally Variable RMET scores from Total Words Uttered, LR3PMS Uttered, and the interaction between Total Words Uttered and LR3PMS Uttered. Model 8 was found to be statistically significant in its overall fit ($R^2 = 0.3336$, $F(3, 173) = 30.37$, $p < .0001$). It was found that Total Words Uttered ($\beta = 0.006686$, $p < .0001$), LR3PMS Uttered ($\beta = 0.0847$, $p = .019$), and their interaction ($\beta = -.00006271$, $p < .0001$) all significantly predicted Unstandardized Culturally Variable RMET scores. Thus, for a single-word increase in the count of Total Words uttered, the best-fit model predicts Unstandardized Culturally Variable RMET performance to increase by approximately 0.007 points. As the RMET is scored on an integer scale ranging from 0 to 36, it is perhaps more meaningful to frame this result in terms of the change in Total Words Uttered expected to correspond to a single-point

increase in Unstandardized Culturally Variable RMET score. Characterized like this, a single-point increase in Unstandardized Culturally Variable RMET score is expected for each increase of 150 words in Total Words Uttered. For a single-word increase in the count of LR3PMS uttered, the best fit model predicts Unstandardized Culturally Variable RMET scores to increase by approximately 0.085 points. As before, this means that for a 12 word increase in the count of LR3PMS uttered, Unstandardized Culturally Variable RMET score is expected to increase by approximately 1 point. Finally, the interaction term can be understood as follows. For a single-word increase in the count of Total Words Uttered, each single-word increase in the count of LR3PMS uttered is expected to *decrease* Unstandardized Culturally Variable RMET scores by 0.000063 points. **Figure S25B** illustrates the predicted Unstandardized Culturally Variable RMET scores with 95% confidence intervals across the range of values for Total Words Uttered. **Figure S26B** illustrates the predicted Unstandardized Culturally Variable RMET scores with 95% confidence intervals across the range of values for LR3PMS. **Figure S27B** illustrates the effect of LR3PMS Uttered on Unstandardized Culturally Variable RMET scores with 95% intervals for participants with a Total Words Uttered count of 395 (the value corresponding to the lower quartile of words uttered), 790 (the median number of words uttered), and 1261 words (the value corresponding to the upper quartile of words uttered). **Figure S28B** is essentially the same as **Figure S27B**, though the x-axis now corresponds to the count of Total Words Uttered. **Figure S28B** illustrates the effect of Total Words Uttered on Unstandardized Culturally Variable RMET scores with 95% intervals for participants with an LR3PMS count of 1 (the value corresponding to the lower quartile of LR3PMS uttered), 3 (the median number of LR3PMS uttered), and 4 words (the value corresponding to the upper quartile of LR3PMS uttered). **Figure S30** contains visualizations of the fit statistics of Model 8.

Model selection and analysis for Standardized Culturally Variable RMET Scores

LR3PMS Uttered (Wellman and Estes Terms). In all 15 models, simple linear regression was used to determine if the predictor variables significantly predicted Standardized Culturally Variable RMET scores. After running all 15 models, Model 8 was found to have the lowest AIC score (AIC: 441.5662) and predicted Standardized Culturally Variable RMET scores from Total Words Uttered, LR3PMS Uttered, and the interaction between Total Words Uttered and LR3PMS Uttered. Model 8 was found to be statistically significant in its overall fit (R^2 = 0.3061, $F(3, 173) = 26.87$, $p < .0001$). It was found that Total Words Uttered ($\beta = 0.0013855$, $p < .0001$), LR3PMS Uttered ($\beta = 0.132486$, $p = .0102$), and their interaction ($\beta = -.000139$, $p < .001$) all significantly predicted Standardized Culturally Variable RMET scores. Thus, for a single-word increase in the count of Total Words uttered, the best-fit model predicts RMET scores to increase by approximately 0.0014 standard deviations. As standardized RMET scores are not intrinsically interpretable on their own, it is perhaps more meaningful to frame this result in terms of the change in Total Words Uttered expected to correspond to a score one standard deviation above the mean. Characterized like this, RMET scores are expected to increase by one standard deviation above the population mean for an increase of 722 words in the count of Total Words Uttered. For a single-word increase in the count of LR3PMS uttered, the best fit model predicts RMET scores to increase by approximately 0.1325 standard deviations. As before, this means that RMET scores are expected to increase by one standard deviation above the population mean for an increase of eight words in the count of LR3PMS uttered. Finally, the interaction term can be understood as follows. For a single-word increase in the count of Total Words Uttered, each single-word increase in the count of LR3PMS uttered is expected to *decrease* RMET scores by 0.00014 standard deviations. **Figure S31A** illustrates the predicted Standardized Culturally Variable RMET scores with 95% confidence intervals across the range of values for Total Words Uttered. **Figure S32A** illustrates the predicted Standardized Culturally Variable RMET scores with 95% confidence intervals across the range of values for LR3PMS.

Figure S33A illustrates the effect of LR3PMS Uttered on Standardized Culturally Variable RMET scores with 95% intervals for participants with a Total Words Uttered count of 395 (the value corresponding to the lower quartile of words uttered), 790 (the median number of words uttered), and 1261 words (the value corresponding to the upper quartile of words uttered). **Figure S34A** is essentially the same as **Figure S33A**, though the x-axis now corresponds to the count of Total Words Uttered. **Figure S34A** illustrates the effect of Total Words Uttered on Standardized Culturally Variable RMET scores with 95% intervals for participants with an LR3PMS count of 1 (the value corresponding to the lower quartile of LR3PMS uttered), 3 (the median number of LR3PMS uttered), and 4 words (the value corresponding to the upper quartile of LR3PMS uttered). **Figure S35** contains visualizations of the fit statistics of Model 8.

LR3PMS Uttered (All Mental State Terms). In all 15 models, simple linear regression was used to determine if the predictor variables significantly predicted Standardized Culturally Variable RMET scores. After running all 15 models, Model 8 was found to have the lowest AIC score (AIC: 433.3395) and predicted Standardized Culturally Variable RMET scores from Total Words Uttered, LR3PMS Uttered, and the interaction between Total Words Uttered and LR3PMS Uttered. Model 8 was found to be statistically significant in its overall fit ($R^2 = 0.3376$, $F(3, 173) = 30.9$, $p < .0001$). It was found that Total Words Uttered ($\beta = 0.001521$, $p < .0001$), LR3PMS Uttered ($\beta = 0.01686$, $p = .0228$), and their interaction ($\beta = -.0000144$, $p < .0001$) all significantly predicted Standardized Culturally Variable RMET scores. Thus, for a single-word increase in the count of Total Words uttered, the best-fit model predicts RMET scores to increase by approximately 0.0015 standard deviations. As standardized RMET scores are not intrinsically interpretable on their own, it is perhaps more meaningful to frame this result in terms of the change in Total Words Uttered expected to correspond to a score one standard deviation above the mean. Characterized like this, RMET scores are expected to increase by one standard deviation above the population mean for an increase of 658 words in the count of Total

Words Uttered. For a single-word increase in the count of LR3PMS uttered, the best fit model predicts RMET scores to increase by approximately 0.01686 standard deviations. As before, this means that RMET scores are expected to increase by one standard deviation above the population mean for an increase of 60 words in the count of LR3PMS uttered. Finally, the interaction term can be understood as follows. For a single-word increase in the count of Total Words Uttered, each single-word increase in the count of LR3PMS uttered is expected to *decrease* RMET scores by 0.000014 standard deviations. **Figure S31B** illustrates the predicted Standardized Culturally Variable RMET scores with 95% confidence intervals across the range of values for Total Words Uttered. **Figure S32B** illustrates the predicted Standardized Culturally Variable RMET scores with 95% confidence intervals across the range of values for LR3PMS. **Figure S33B** illustrates the effect of LR3PMS Uttered on Standardized Culturally Variable RMET scores with 95% intervals for participants with a Total Words Uttered count of 395 (the value corresponding to the lower quartile of words uttered), 790 (the median number of words uttered), and 1261 words (the value corresponding to the upper quartile of words uttered). **Figure S34B** is essentially the same as **Figure S33B**, though the x-axis now corresponds to the count of Total Words Uttered. **Figure S34B** illustrates the effect of Total Words Uttered on Standardized Culturally Variable RMET scores with 95% intervals for participants with an LR3PMS count of 1 (the value corresponding to the lower quartile of LR3PMS uttered), 3 (the median number of LR3PMS uttered), and 4 words (the value corresponding to the upper quartile of LR3PMS uttered). **Figure S36** contains visualizations of the fit statistics of Model 8.

Table S1

Variance Estimates and Standard Deviations of Variance Component Model 2 Random Effects

Variable	Random Effect	Variance	Std.Dev.	Groups
Participant ID	Intercept	0.12428	0.3525	177
Interaction between Field Site and Video ID	Intercept	0.56473	0.7515	27
Video ID	Intercept	0.4947	0.7034	9
Field Site	Intercept	0.02472	0.1572	3
Number of observations: 1589				

Note. Variance estimates for each of the random effects included in the model vary substantially in the amount of variance attributed to each. Estimates are reported untransformed and thus represent the variance in the log of the counts of LR3PMS.

Table S2

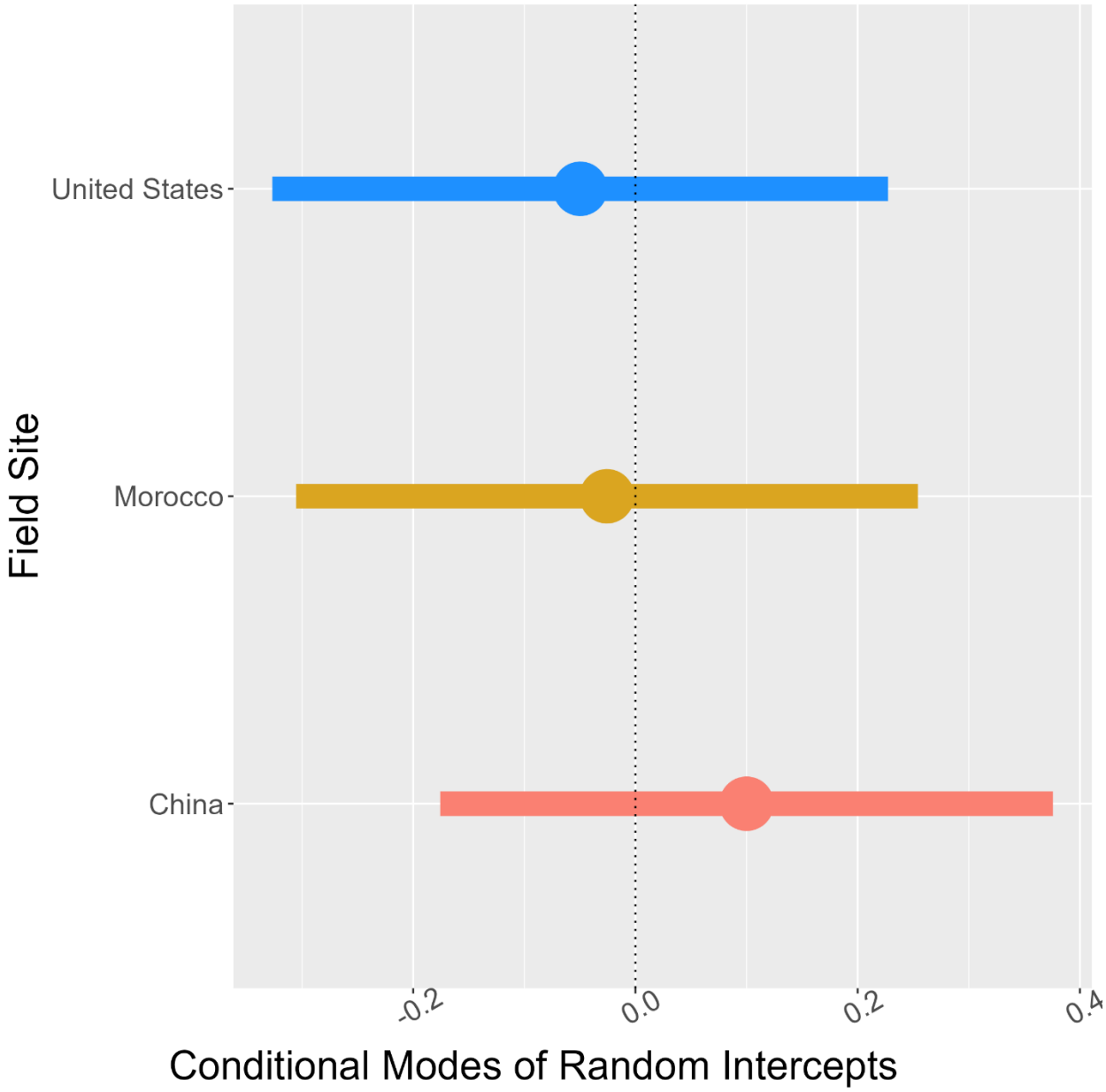
Variance Estimates and Standard Deviations of Variance Component Model 2 Random Effects

Variable	Random Effect	Variance	Std.Dev.	Groups
Participant ID	Intercept	0.03701	0.1924	177
Interaction between Field Site and Video ID	Intercept	0.05245	0.2290	27
Video ID	Intercept	0.0915	0.3025	9
Field Site	Intercept	0.0000000006461	0.0000254	3
Number of observations: 1589				

Note. Variance estimates for each of the random effects included in the model vary substantially in the amount of variance attributed to each. Estimates are reported untransformed and thus represent the variance in the log of the counts of LR3PMS.

Figure S1

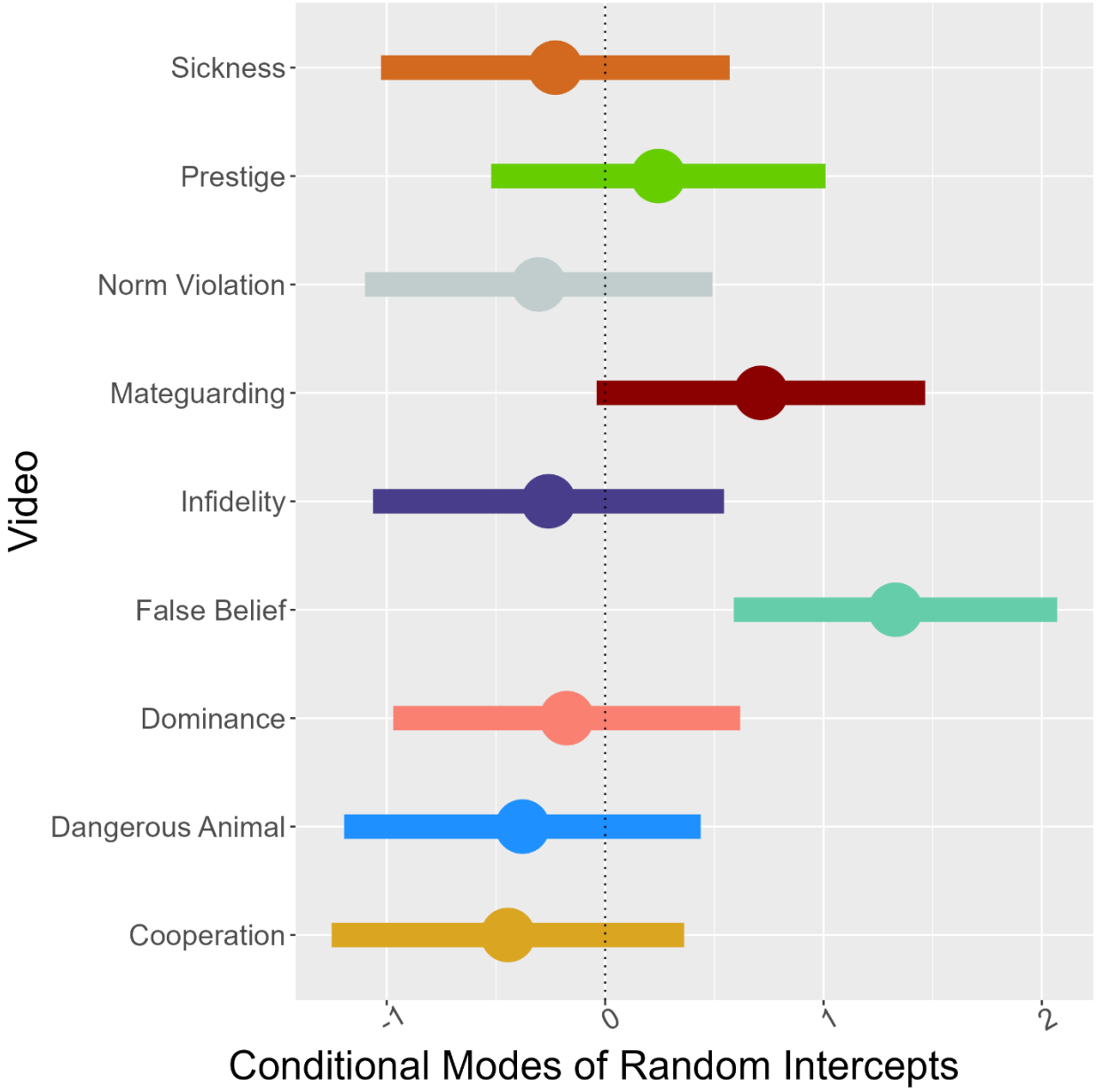
Conditional Modal Estimates of Random Intercepts for Levels of 'Field Site' Factor in VCM 2



Note. Conditional modal estimates of the random intercepts for each level of the 'Field Site' factor with 95% confidence intervals derived from VCM 2. Units are untransformed and therefore represent the difference between the log of the expected count of LR3PMS overall and the log of the expected count of LR3PMS for each level, holding all other variables constant.

Figure S2

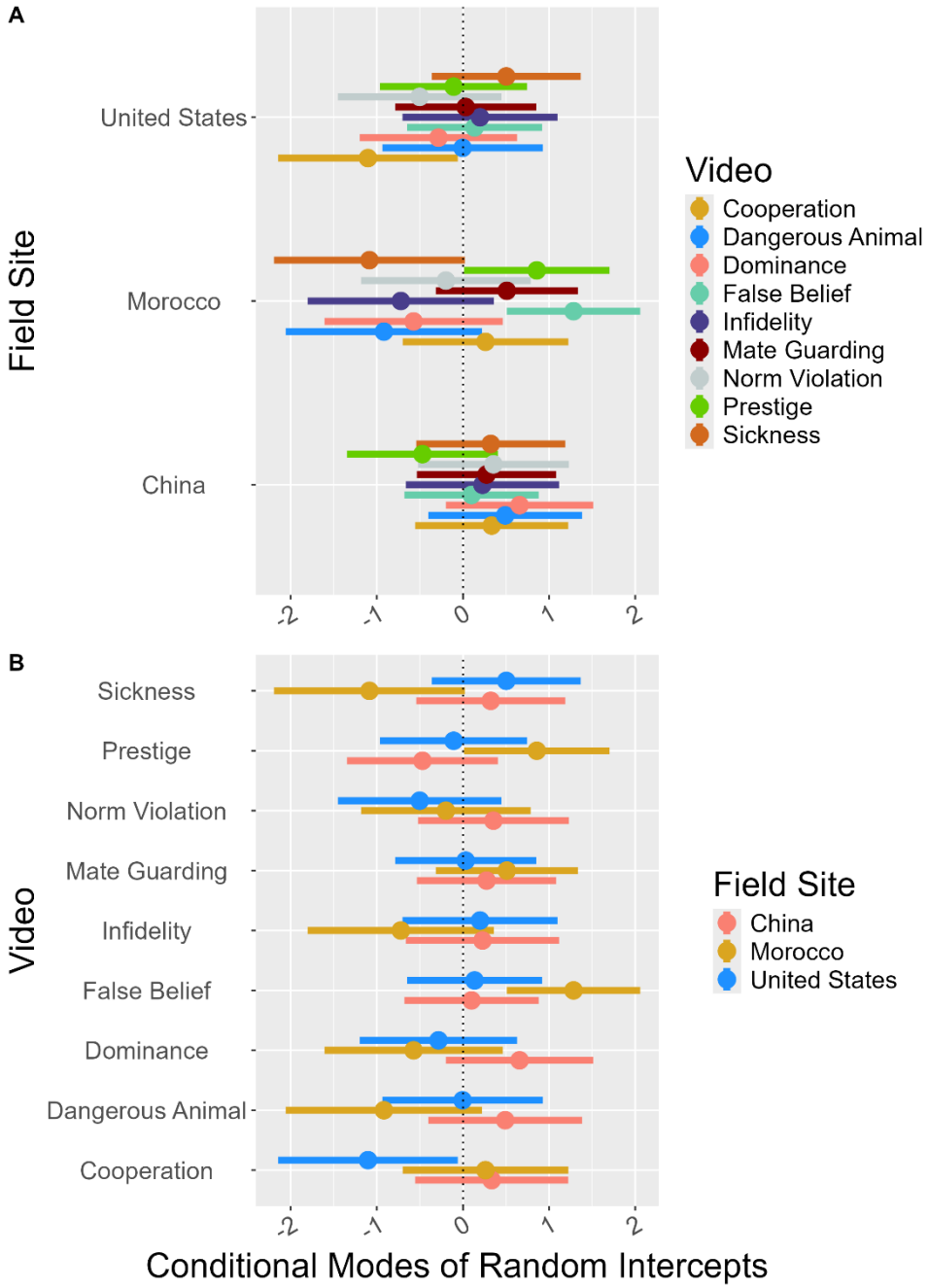
Conditional Modal Estimates of Random Intercepts for Levels of 'Video ID' Factor in VCM 2



Note. Conditional modal estimates of the random intercepts for each level of the 'Video ID' factor with 95% confidence intervals derived from VCM 2. Units are untransformed and therefore represent the difference between the log of the expected count of LR3PMS overall and the log of the expected count of LR3PMS for each level, holding all other variables constant.

Figure S3

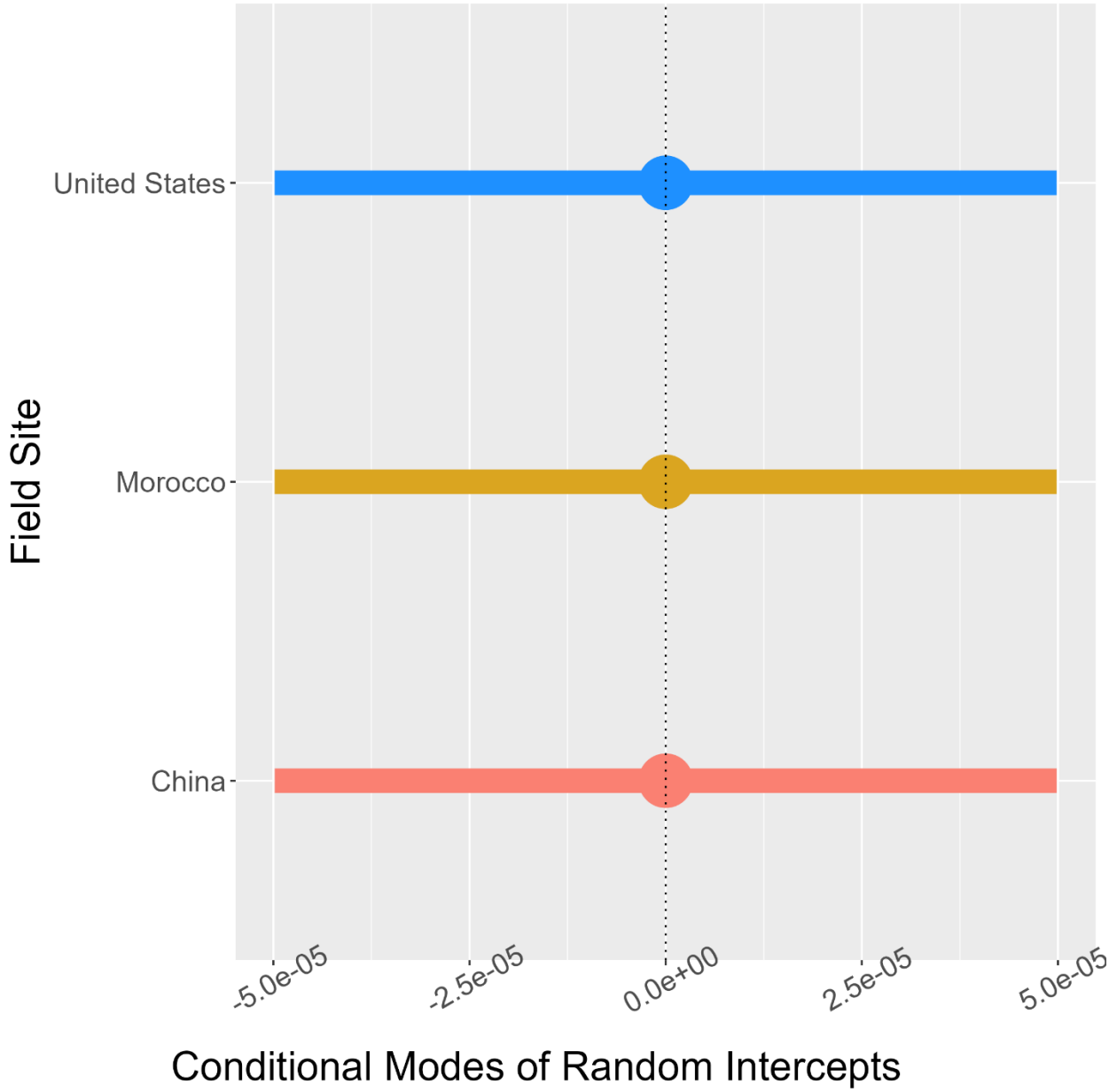
Conditional Modal Estimates of Random Intercepts for Levels of 'Video ID by Field Site' Interaction Factor Grouped by 'Field Site' Factor in VCM 2



Note. Conditional modal estimates of the random intercepts for each level of the VCM 2 interaction factor with 95% confidence intervals. Units are untransformed and therefore represent the difference between the log of the expected count of LR3PMS overall and the log of the expected count of LR3PMS for each level, holding all other variables constant. (A) Estimates for *Video ID* are grouped by *Field Site* on the Y axis. (B) Estimates *Field Site* are grouped by *Video ID* on the y-axis.

Figure S4

Conditional Modal Estimates of Random Intercepts for Levels of 'Field Site' Factor in VCM 2



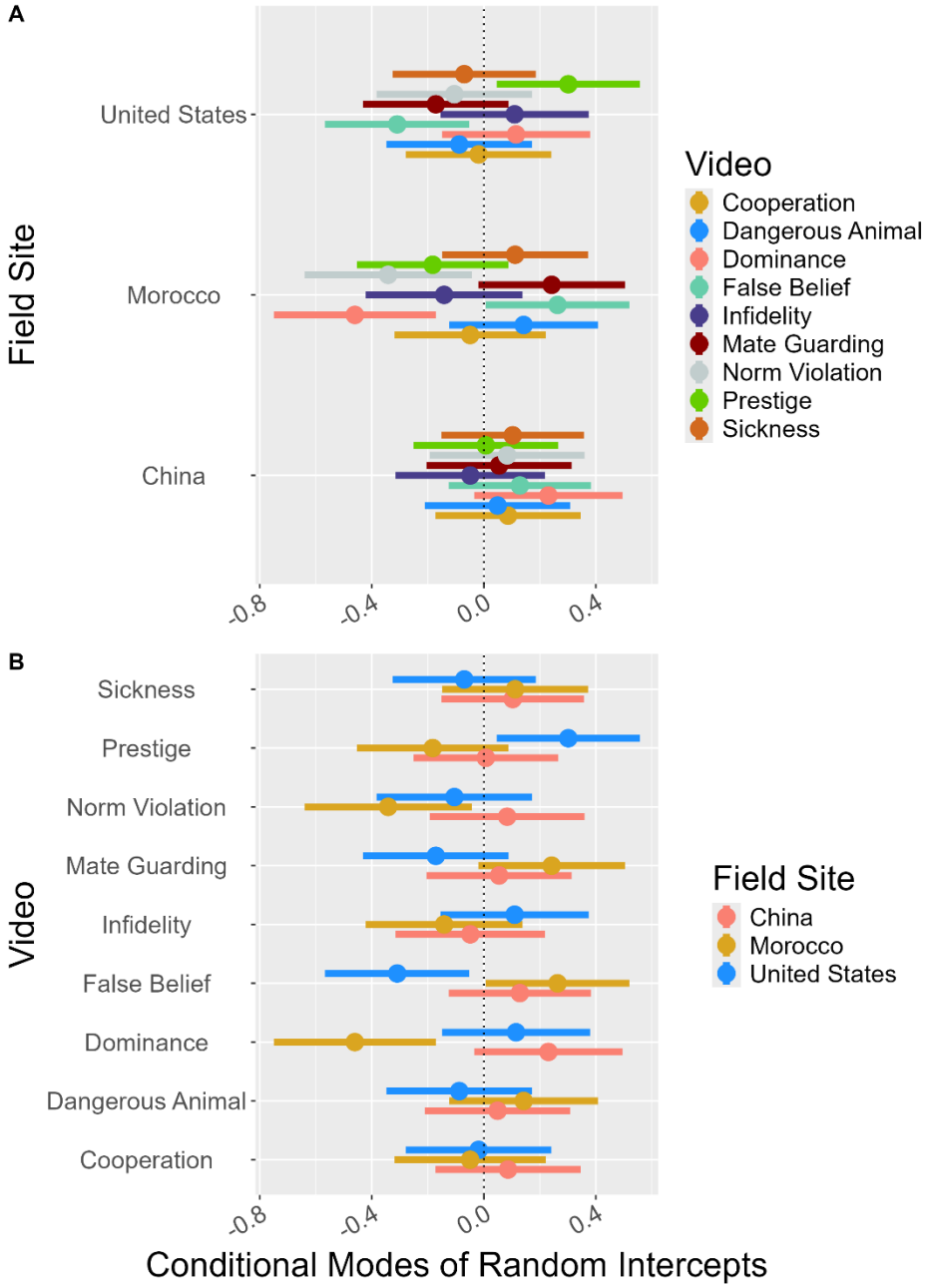
Note. Conditional modal estimates of the random intercepts for each level of the 'Field Site' factor with 95% confidence intervals derived from VCM 2. Units are untransformed and therefore represent the difference between the log of the expected count of LR3PMS overall and the log of the expected count of LR3PMS for each level, holding all other variables constant.

Figure S5
Conditional Modal Estimates of Random Intercepts for Levels of 'Video ID' Factor in VCM 2

Note. Conditional modal estimates of the random intercepts for each level of the 'Video ID' factor with 95% confidence intervals derived from VCM 2. Units are untransformed and therefore represent the difference between the log of the expected count of LR3PMS overall and the log of the expected count of LR3PMS for each level, holding all other variables constant.

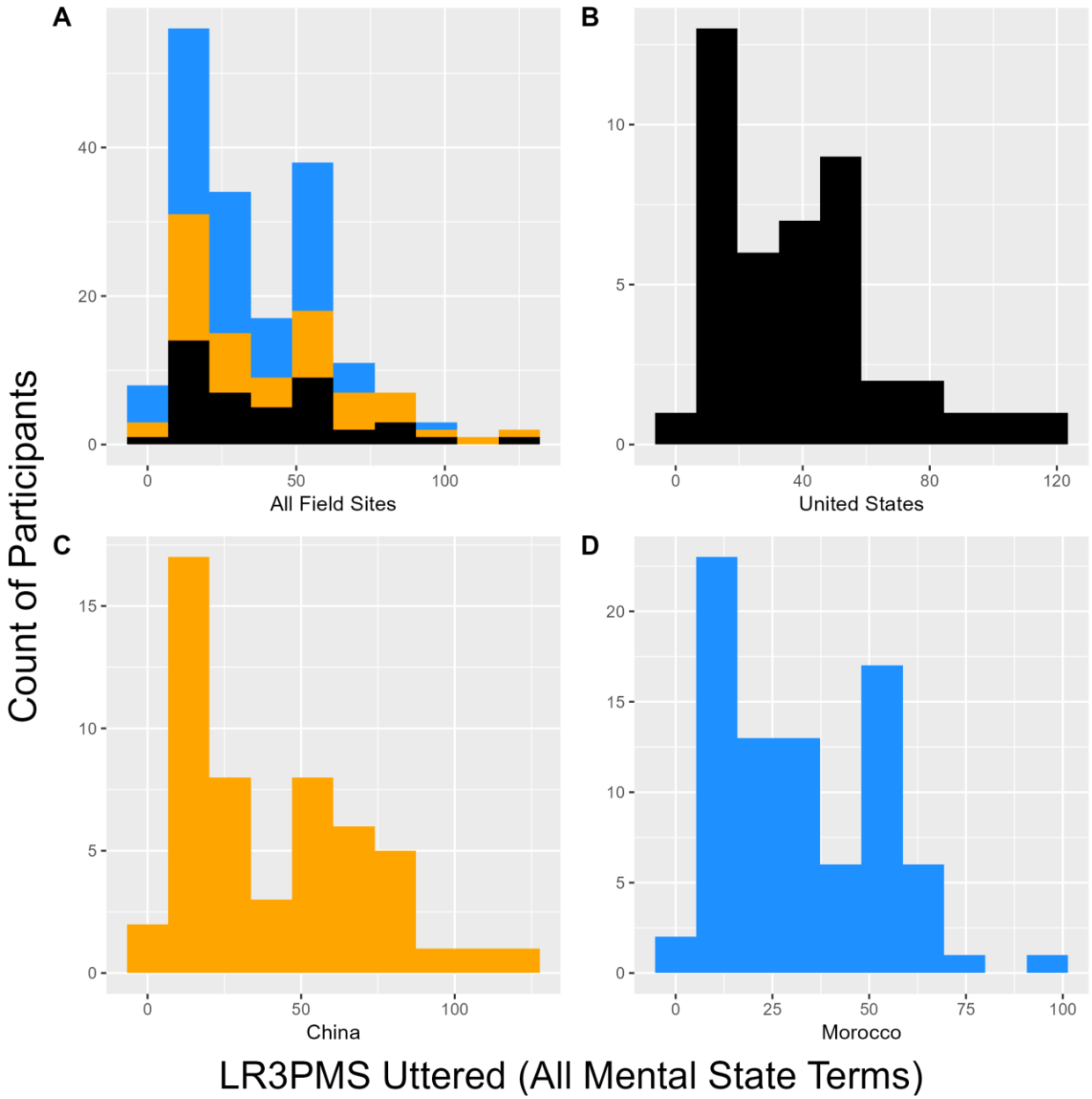
Figure S7

Conditional Modal Estimates of Random Intercepts for Levels of 'Video ID by Field Site' Interaction Factor Grouped by 'Field Site' Factor in VCM 2



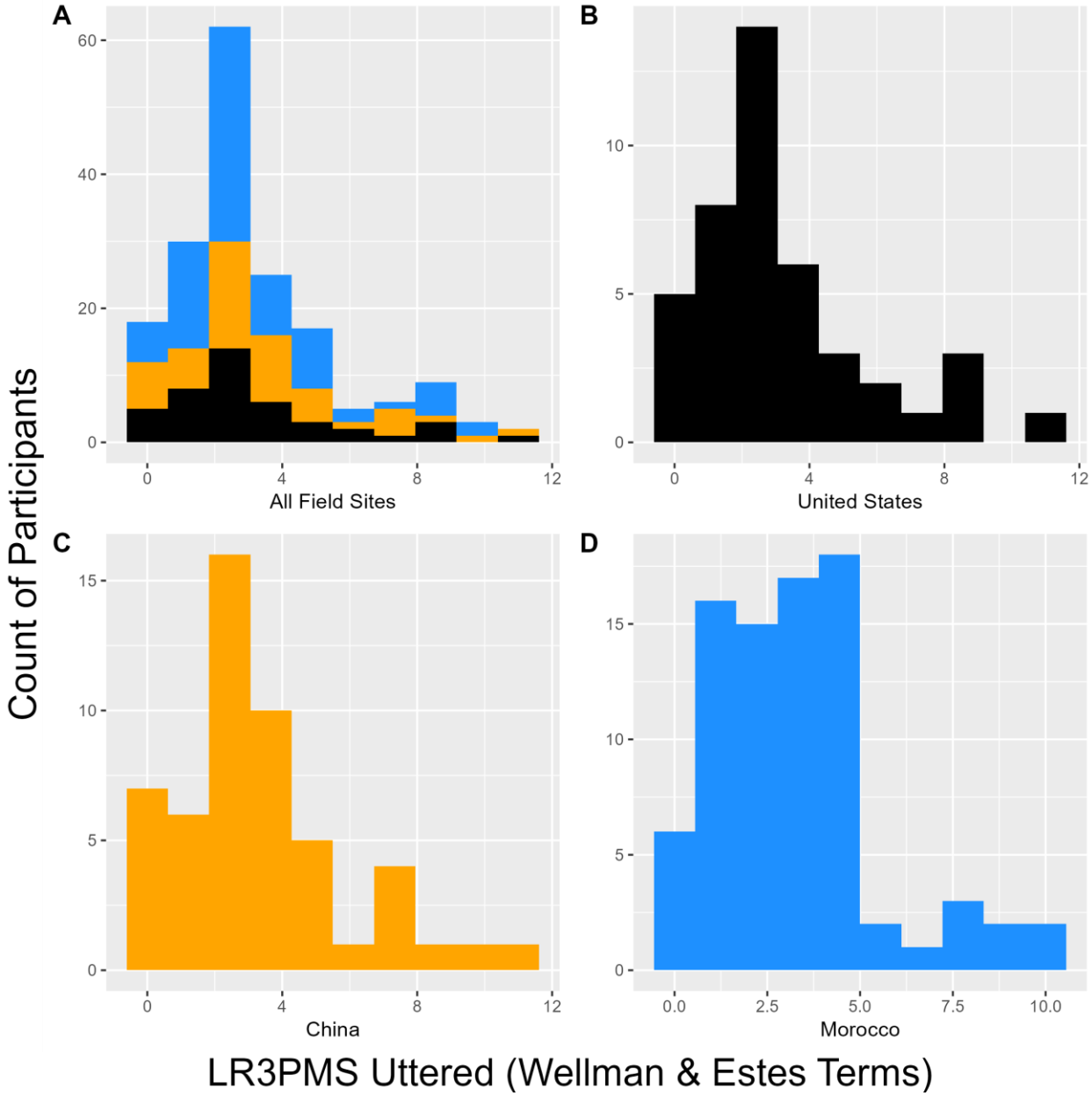
Note. Conditional modal estimates of the random intercepts for each level of the VCM 2 interaction factor with 95% confidence intervals. Units are untransformed and therefore represent the difference between the log of the expected count of LR3PMS overall and the log of the expected count of LR3PMS for each level, holding all other variables constant. (A) Estimates for *Video ID* are grouped by *Field Site* on the Y axis. (B) Estimates *Field Site* are grouped by *Video ID* on the y-axis.

Figure S8



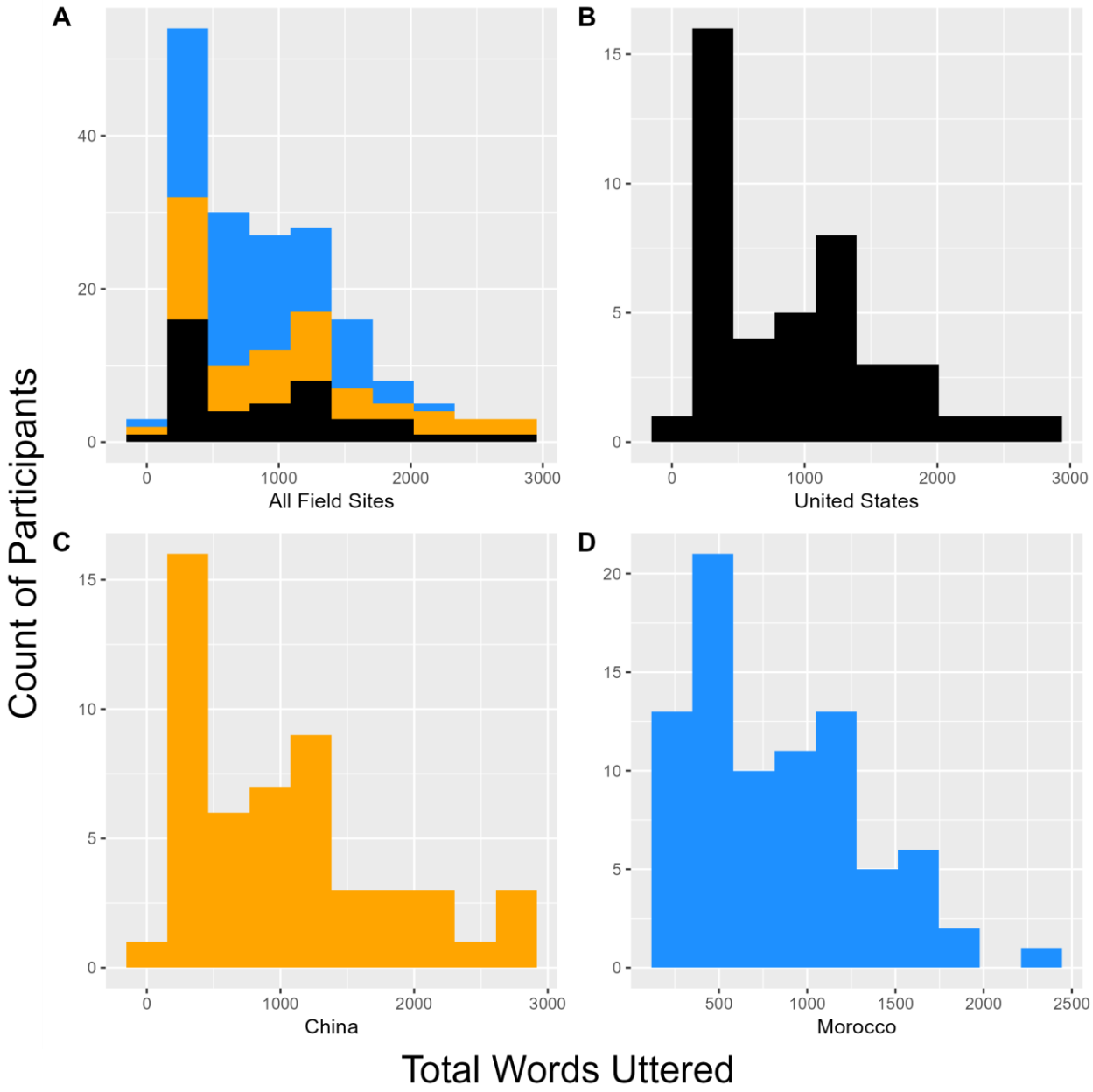
Note. Probability density estimates for the total count of lexical references to third party mental states uttered by each participant. As can be seen, the distributions presented here exhibit right-skewness and exhibit a shape broadly similar to that of a Poisson distribution, as might be expected of count data. The bimodality in the shape of these distributions is unusual and may require further investigation. (S8A) Probability density estimate for all participants with field site from which participants were recruited indicated. Probability density estimates for the total count of lexical references to third party mental states uttered by US participants only, Chinese participants only, and Moroccan participants only can be found in (S8B), (S8C), and (S8D) respectively.

Figure S9



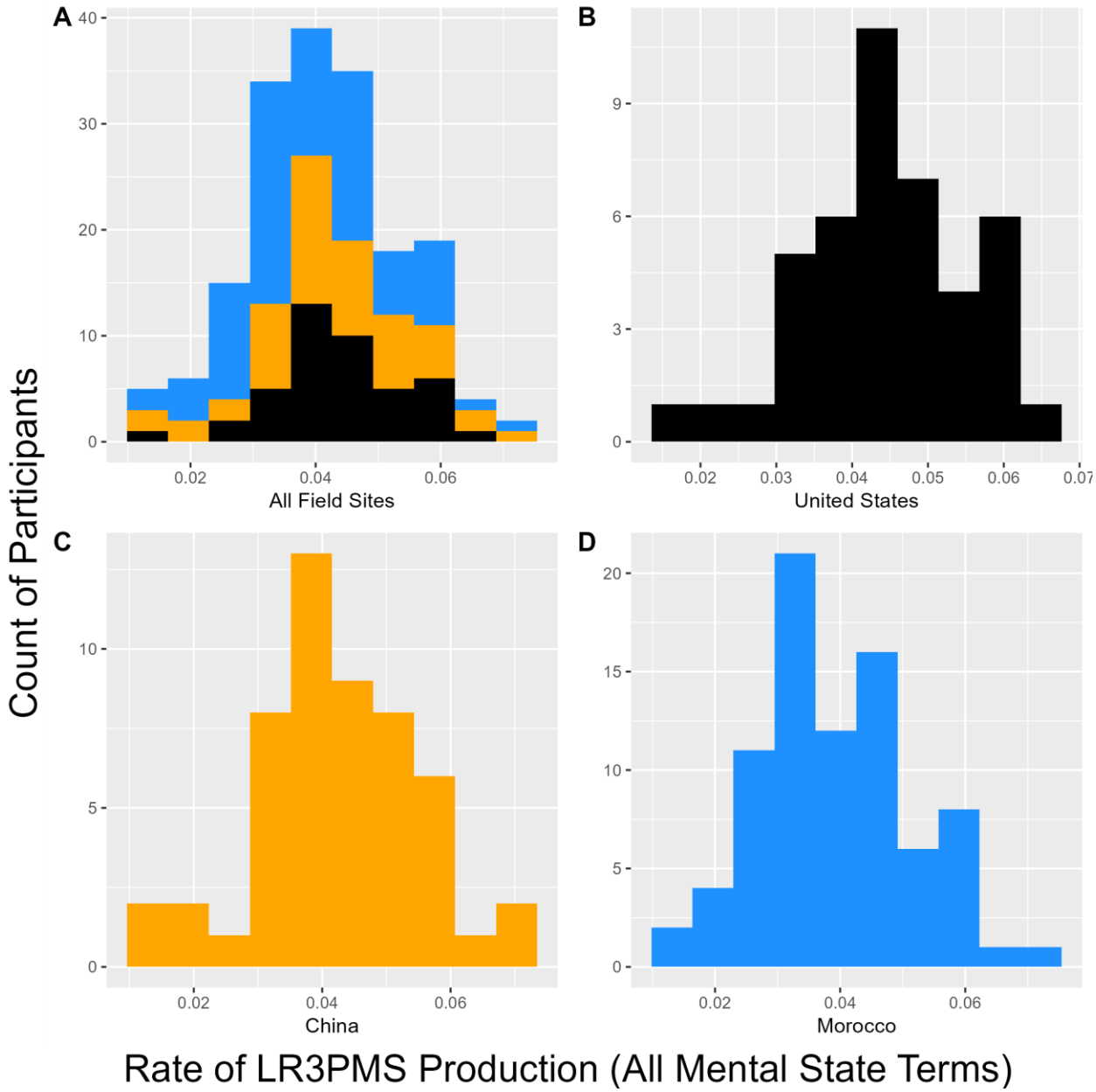
Note. Probability density estimates for the total count of lexical references to third party mental states uttered by each participant. As can be seen, the distributions presented here exhibit right-skewness and exhibit a shape broadly similar to that of a Poisson distribution, as might be expected of count data. The bimodality in the shape of these distributions is unusual and may require further investigation. (S9A) Probability density estimate for all participants with field site from which participants were recruited indicated. Probability density estimates for the total count of lexical references to third party mental states uttered by US participants only, Chinese participants only, and Moroccan participants only can be found in (S9B), (S9C), and (S9D) respectively.

Figure S10



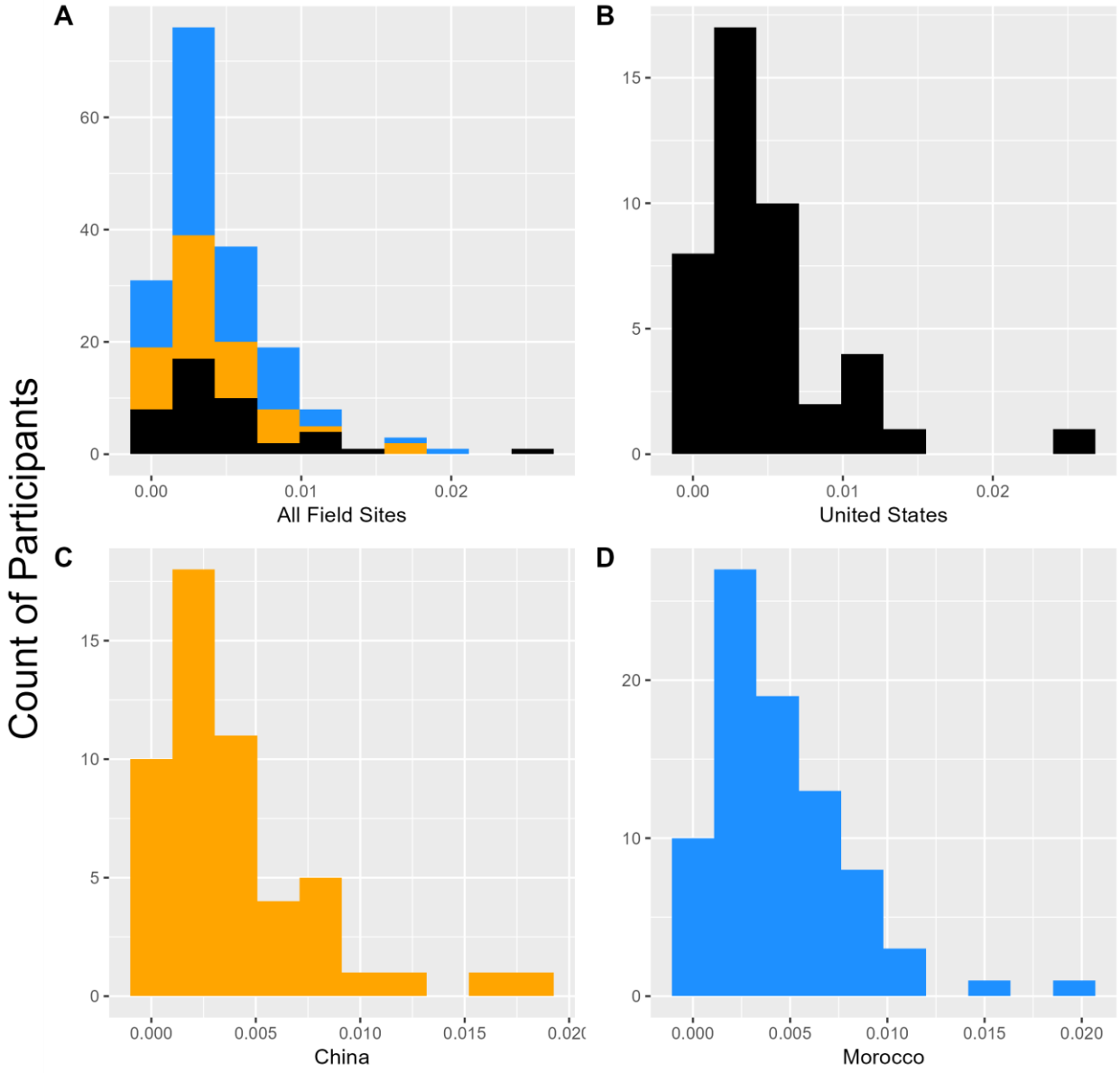
Note. Probability density estimates for the total count of words uttered by each participant overall. As can be seen, the distributions presented here exhibit right-skewness and exhibit a shape broadly similar to that of a Poisson distribution, as might be expected of count data. The bimodality in the shape of these distributions is unusual and may require further investigation. (S10A) Probability density estimate for all participants with field site from which participants were recruited indicated. Probability density estimates for the total count of words uttered by each participant from the United States only, China only, and Morocco only can be found in (S10B), (S10C), and (S10D), respectively.

Figure S11



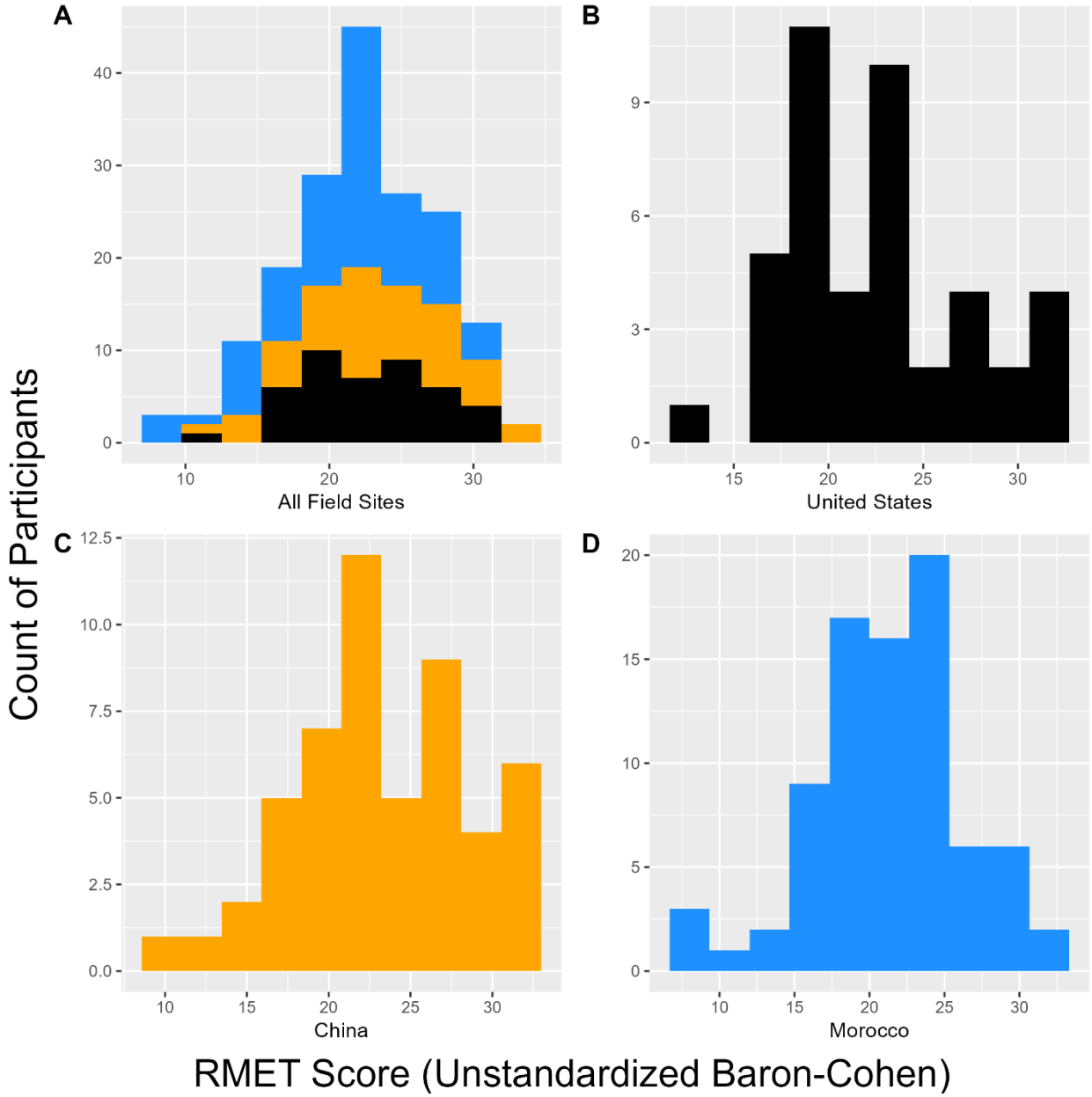
Note. Probability density estimates for the overall rate at which participants produced lexical references to third party mental states. Rates were calculated by dividing the number of lexical references to third party mental states by the total number of words uttered by the participant. Notably, these distributions exhibit a shape that differs from those of each constituent variable (See Figs S8, S9, and S10). This shape is approximately normal, in contrast to the approximately Poisson-shaped distributions of each of the count variables from which the rate was derived. (S11A) Probability density estimate for all participants with field site from which participants were recruited indicated. Probability density estimates for the total count of words uttered by each participant from the United States only, China only, and Morocco only can be found in (S11B), (S11C), and (S11D), respectively.

Figure S12



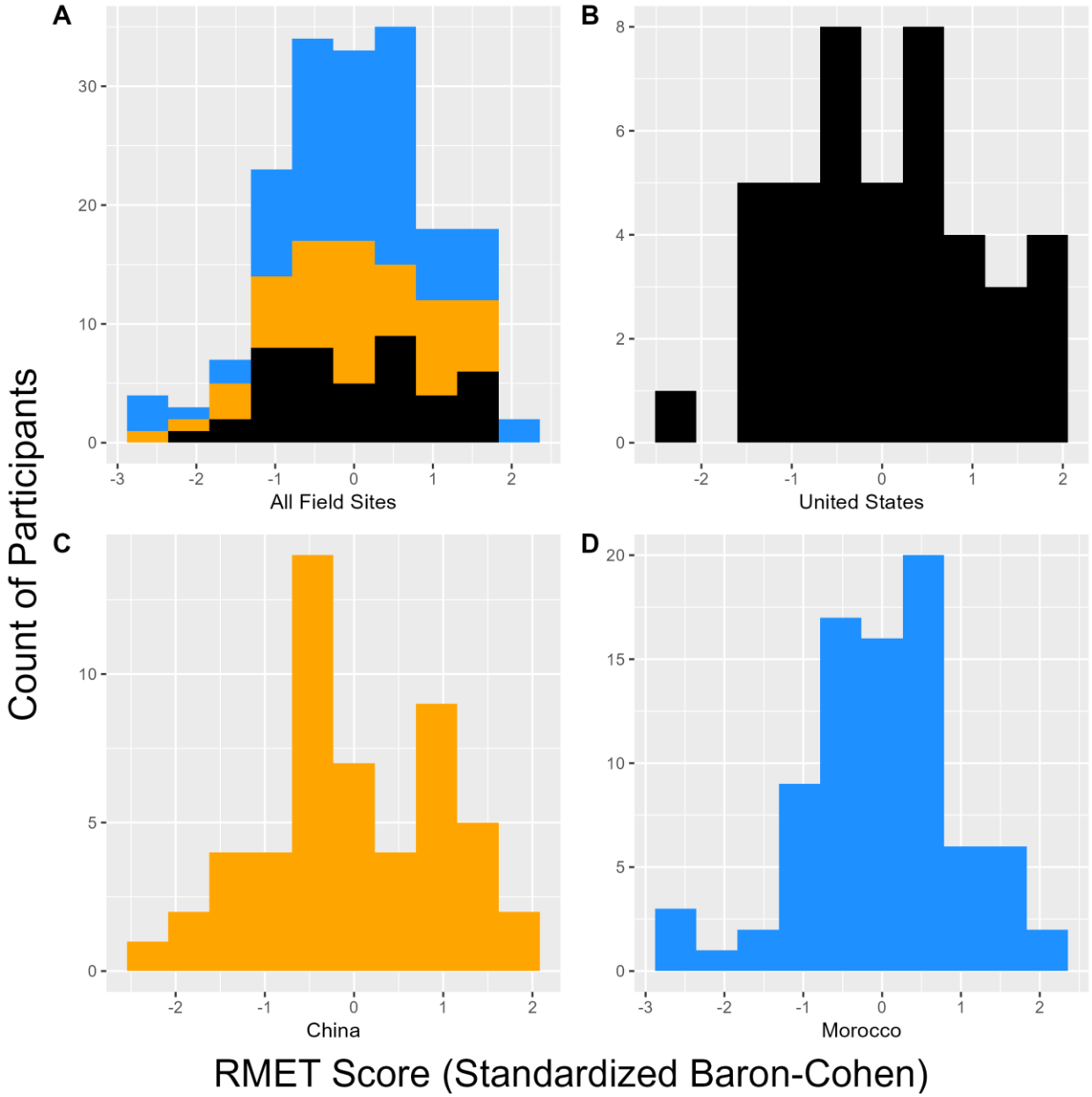
Note. Probability density estimates for the overall rate at which participants produced lexical references to third party mental states. Rates were calculated by dividing the number of lexical references to third party mental states by the total number of words uttered by the participant. Notably, these distributions exhibit a shape that does not differ from those of each constituent variable (See Figs S8, S9, and S10). Both the distributions here and the distributions in Figures S8, S9, and S10 are all approximately Poisson-shaped. (S12A) Probability density estimate for all participants with field site from which participants were recruited indicated. Probability density estimates for the total count of words uttered by each participant from the United States only, China only, and Morocco only can be found in (S12B), (S12C), and (S12D), respectively.

Figure S13



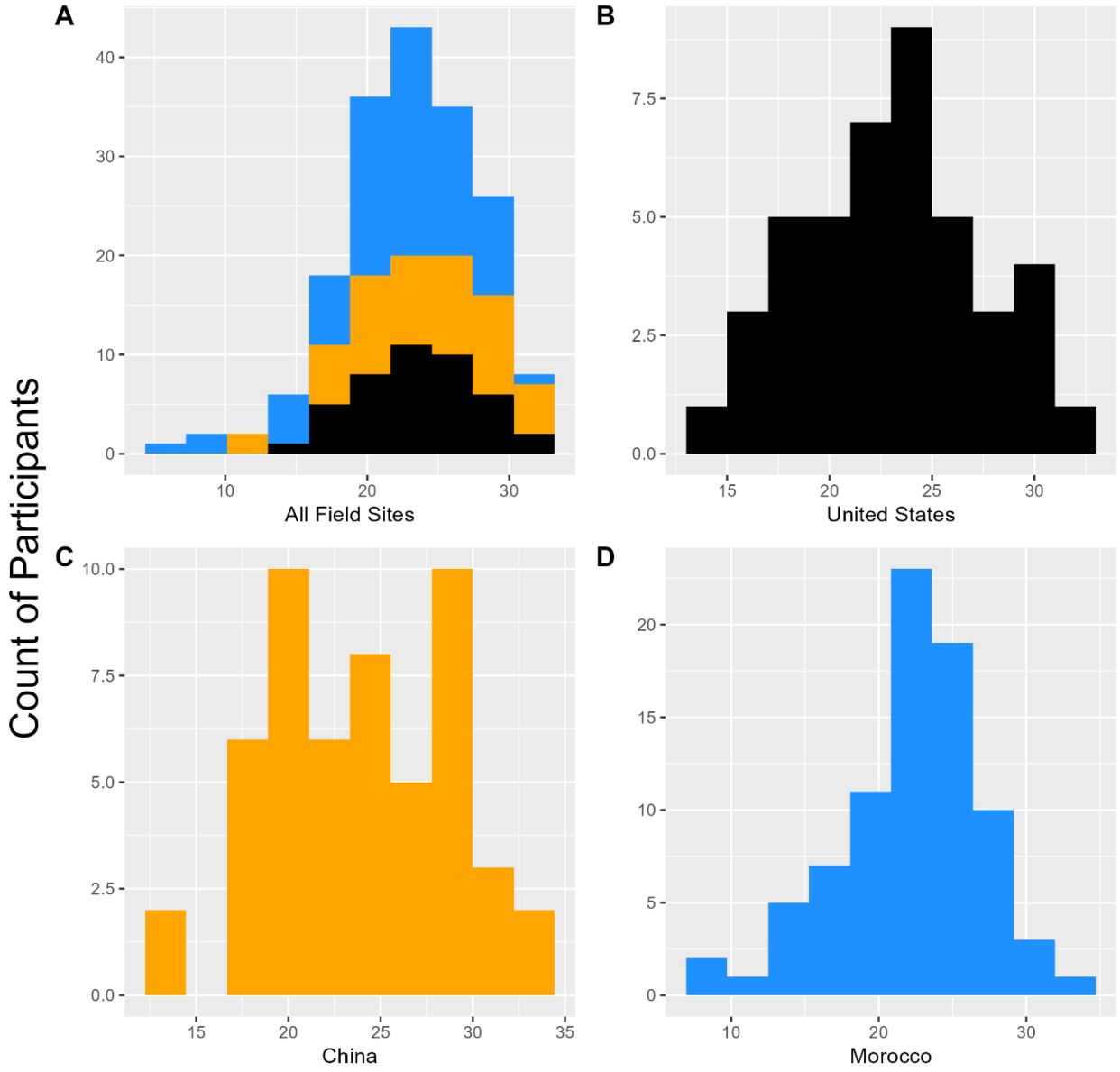
Note. Probability density estimates of participant scores on the RMET using the original, unstandardized scoring procedure as developed by Baron-Cohen et al., (2001). Despite mild negative skewness and some apparent potential bimodality, the shape of the distribution is approximately normal, as confirmed by a Kolmogorov-Smirnov test, D = 0.068, p = 0.3863. (S13A) Probability density estimate for all participants with field site from which participants were recruited indicated. Probability density estimates for participant RMET performance using the Unstandardized Baron-Cohen coding scheme from the United States only, China only, and Morocco only can be found in (S13B), (S13C), and (S13D), respectively.

Figure S14



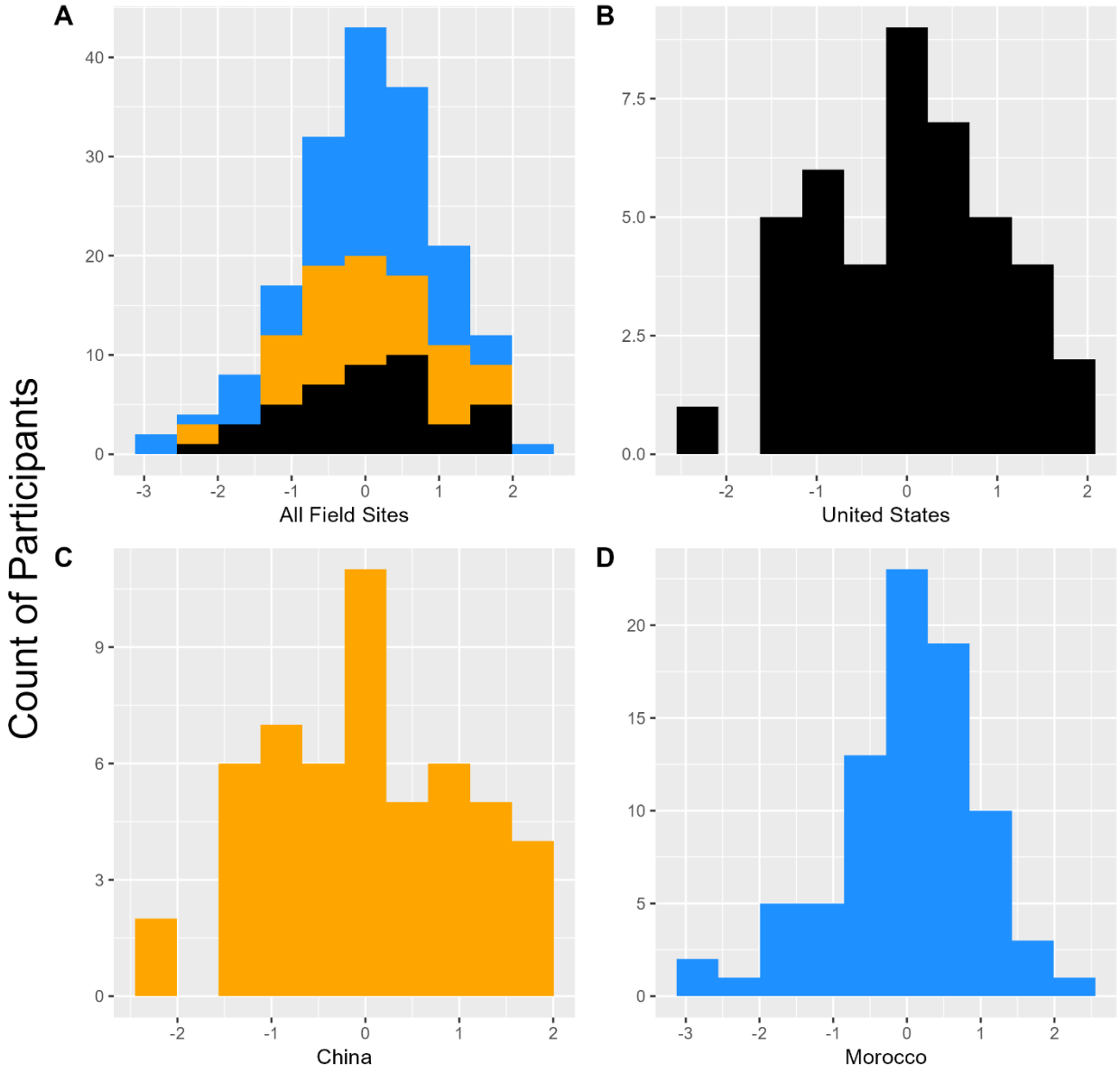
Note. Probability density estimates of participant scores on the RMET using within-field site standardization of participant scores generated using the original coding scheme as developed by Baron-Cohen et al., (2001). Despite mild negative skewness and some apparent potential bimodality, the shape of the distribution is approximately normal, as confirmed by a Kolmogorov-Smirnov test, D = 0.058049, p = 0.5898. (S14A) Probability density estimate for all participants with field site from which participants were recruited indicated. Probability density estimates for participant RMET performance using the Standardized Baron-Cohen coding scheme from the United States only, China only, and Morocco only can be found in (S14B), (S14C), and (S14D), respectively.

Figure S15



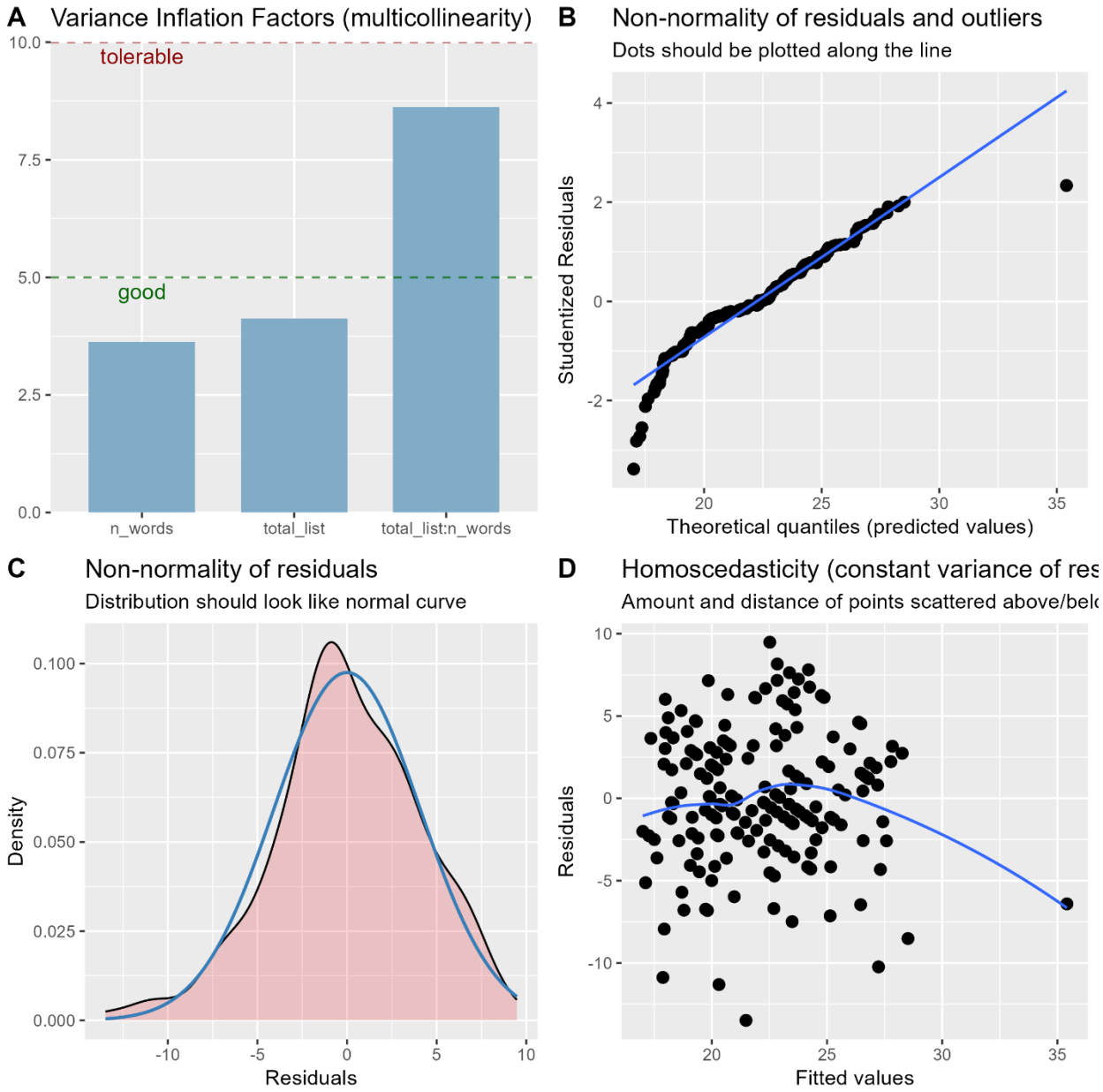
Note. Probability density estimates of participant scores on the RMET generated using a novel procedure to create culturally variable coding schemes based on participant consensus for each item within the field sites sampled. Despite mild negative skewness, the shape of the distribution is approximately normal, as confirmed by a Kolmogorov-Smirnov test, D = 0.042879, p = 0.9008. (S15A) Probability density estimate for all participants with field site from which participants were recruited indicated. Probability density estimates for participant RMET performance using the Unstandardized Culturally Variable coding schemes from the United States only, China only, and Morocco only can be found in (S15B), (S15C), and (S15D), respectively.

Figure S16



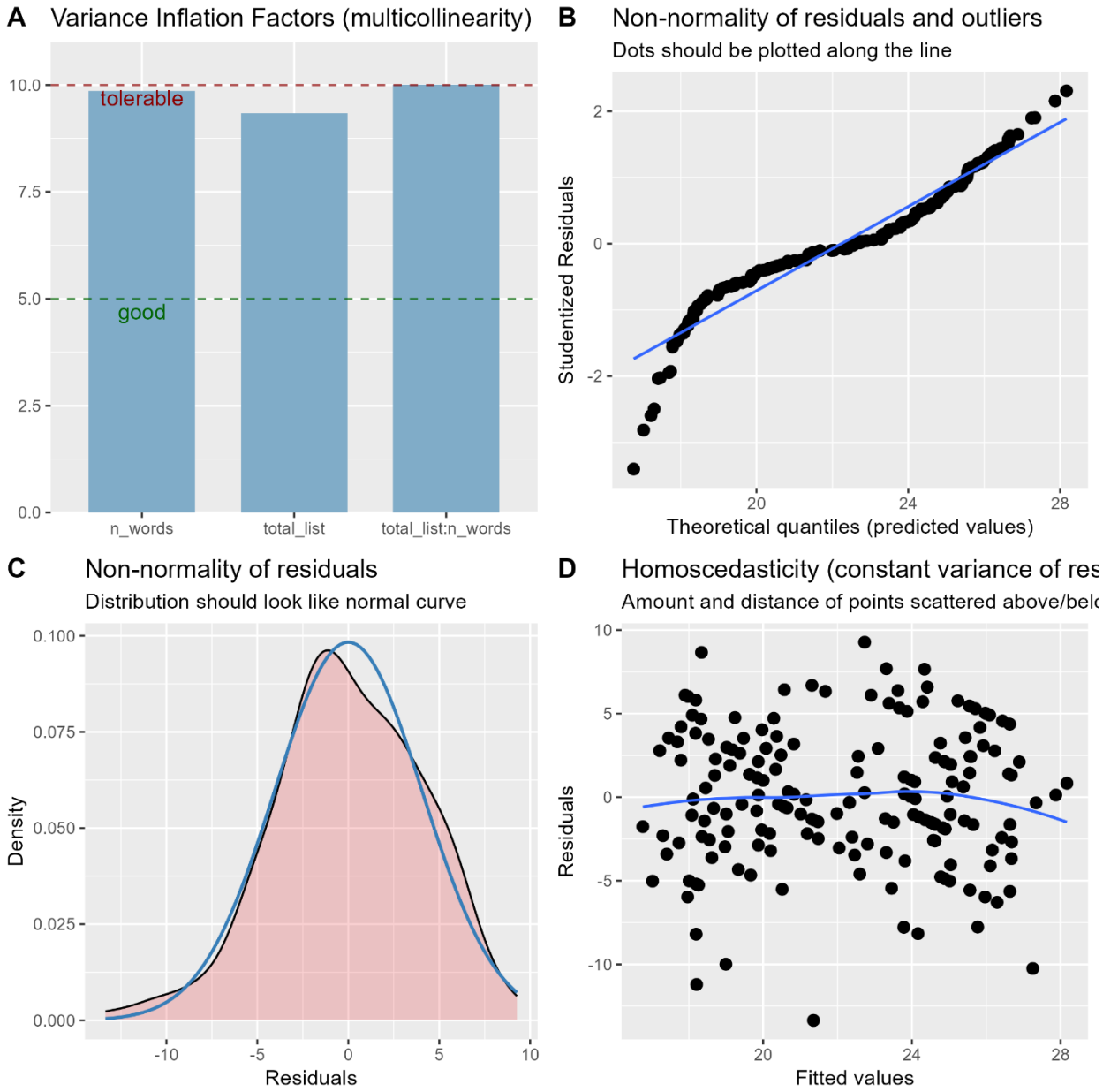
Note. Probability density estimates of within-field site standardized participant scores on the RMET generated using a novel procedure to create culturally variable coding schemes based on participant consensus for each item within the field sites sampled. Despite mild negative skewness, the shape of the distribution is approximately normal, as confirmed by a Kolmogorov-Smirnov test, D = 0.055301, p = 0.6512. (S16A) Probability density estimate for all participants with field site from which participants were recruited indicated. Probability density estimates for participant RMET performance using the Standardized Culturally Variable coding schemes from the United States only, China only, and Morocco only can be found in (S16B), (S16C), and (S16D), respectively.

Figure S17



Note: Fit statistics for Model 8 on participant RMET performance using the Unstandardized Baron-Cohen coding scheme where LR3PMS were coded using the Wellman and Estes Terms coding scheme. (S17A) Variance inflation factors for each of the predictors in Model 8 show that they are uncorrelated. (S17B) QQ plot of model residuals shows that with the exception of some of the lower theoretical quantiles, residuals are normally distributed. (S17C) Another illustration of the residuals showing a normal distribution. (S17D) Variance in the residuals in more or less constant across the range of fitted values.

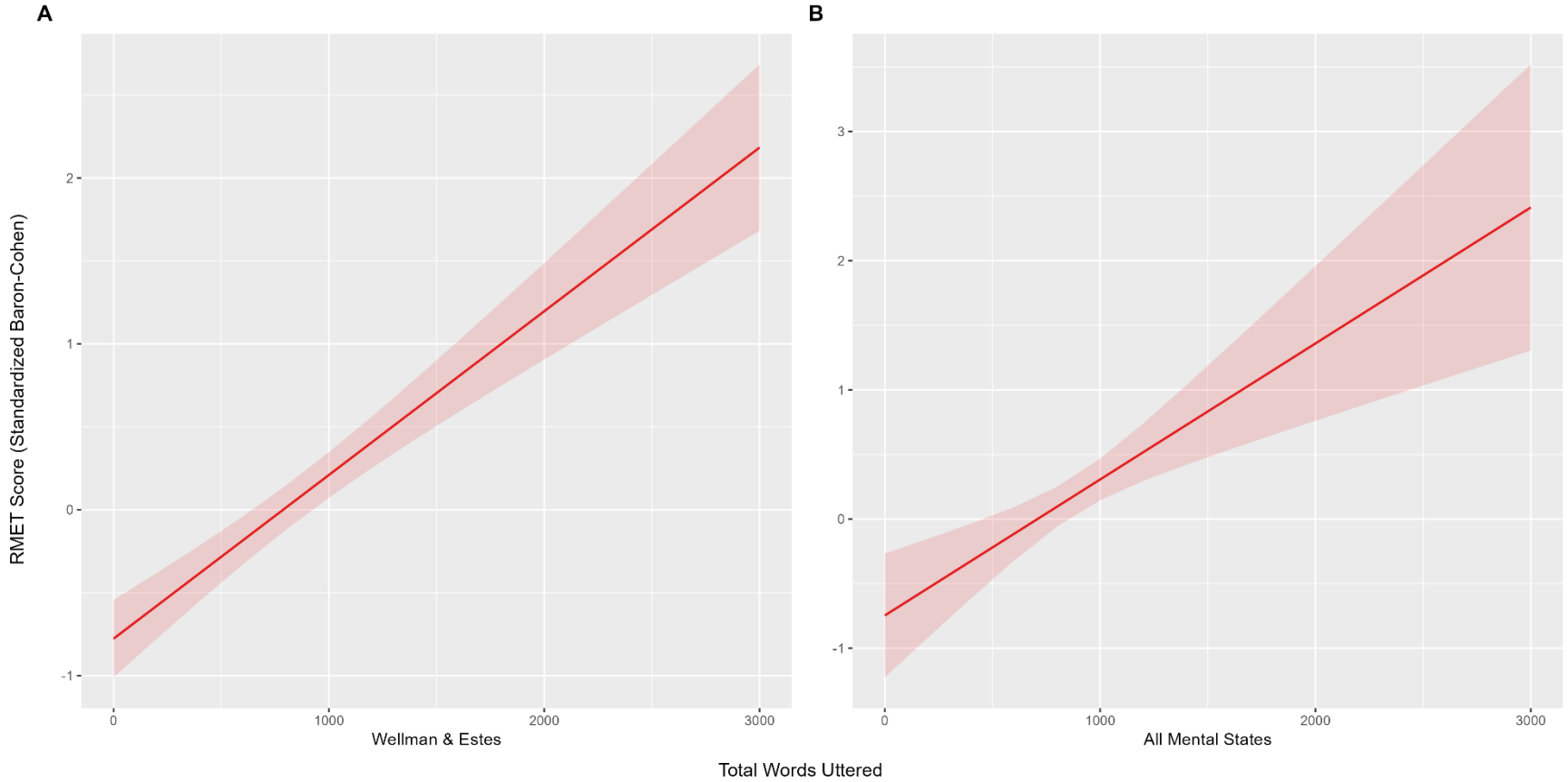
Figure S18



Note: Fit statistics for Model 8 on participant RMET performance using the Unstandardized Baron-Cohen coding scheme where LR3PMS were coded using the All Mental State Terms coding scheme. (S18A) Variance inflation factors for each of the predictors in Model 8 show that they are uncorrelated. (S18B) QQ plot of model residuals shows that with the exception of some of the lower theoretical quantiles, residuals are normally distributed. (S18C) Another illustration of the residuals showing a normal distribution. (S18D) Variance in the residuals in more or less constant across the range of fitted values.

Figure S19

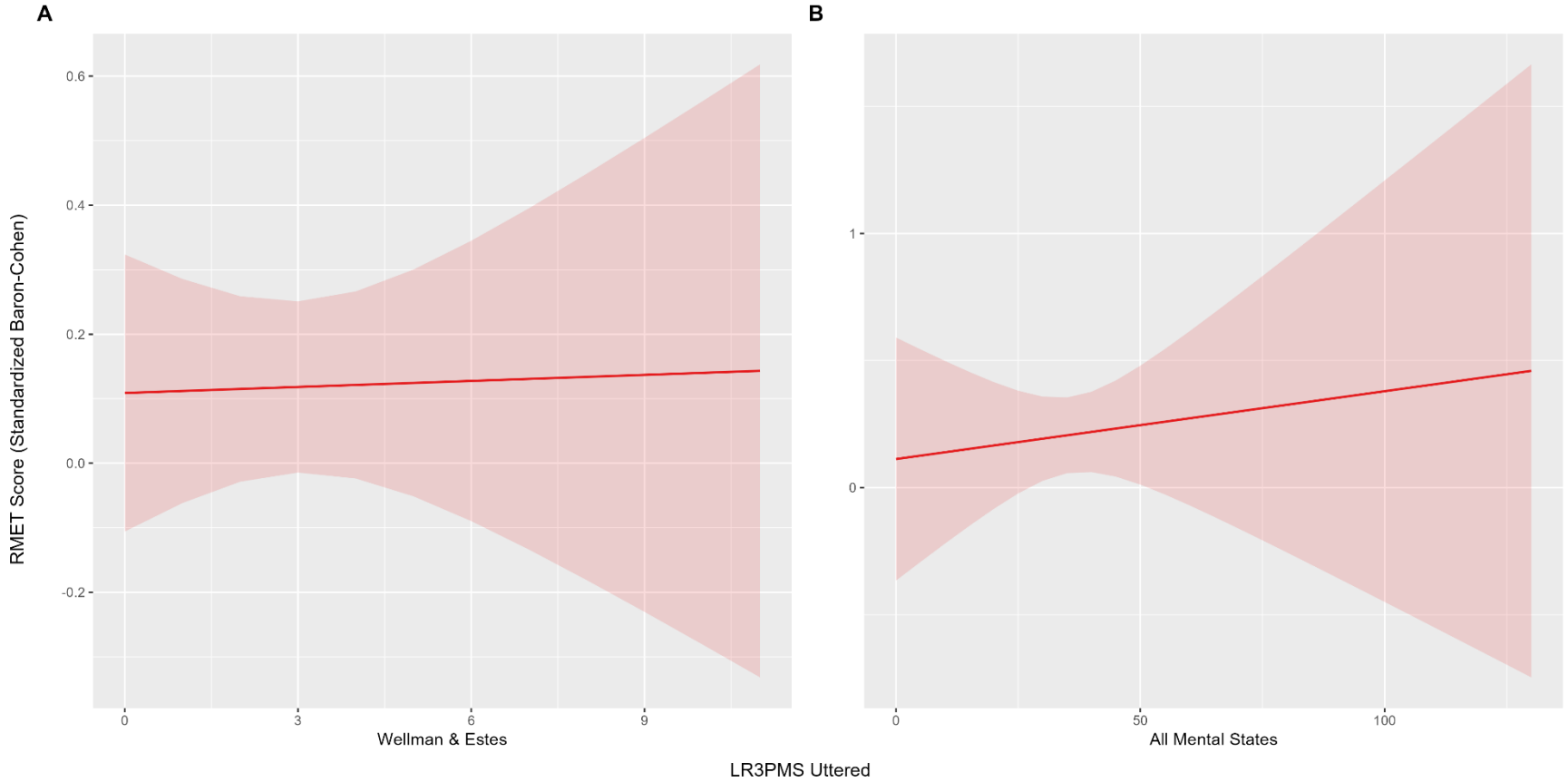
Total Words Uttered by Participants Strongly Positively Predicts Performance on RMET



Note. The predicted results of Model 8 suggest that as the counts of Total Words Uttered increase, so too do participant scores on the Reading the Mind in the Eyes Test (RMET) using the Standardized Baron-Cohen coding scheme. Visualization holds LR3PMS Uttered constant at the sample mean value. (A) Model 8 predictions when using LR3PMS coded using Wellman and Estes terms. (B). Model 8 predictions when using LR3PMS coded using All Mental State terms.

Figure S20

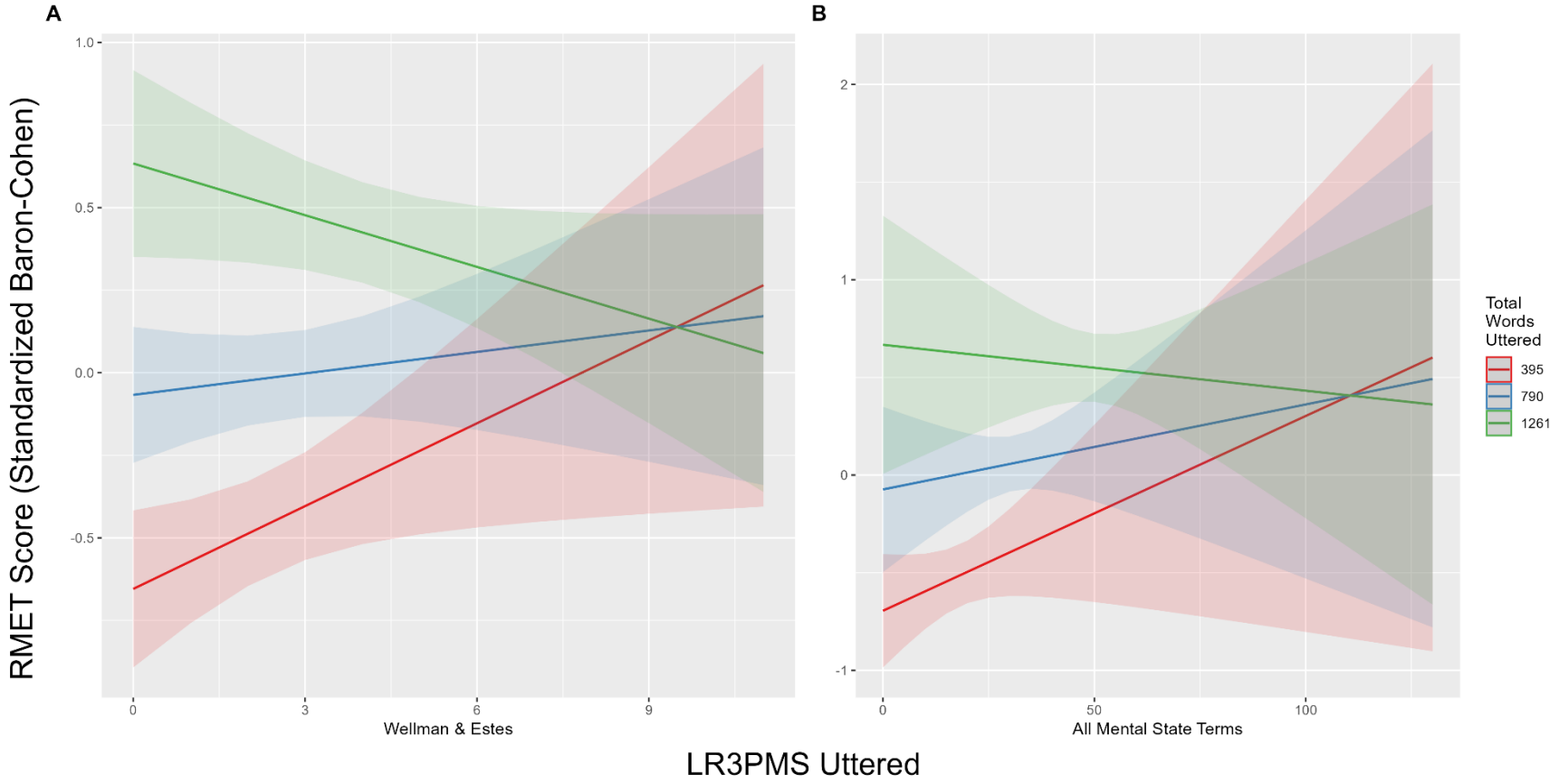
LR3PMS Uttered by Participants Predicts Performance on RMET Less Strongly Than Total Words Uttered



Note. Model 8 predicts that the participants' scores on the RMET will increase modestly as the total number of All Mental State Terms LR3PMS uttered increases. This effect is independent of, albeit weaker than that of Total Words Uttered. Visualizations hold Total Words Uttered constant at the sample mean value. (A) Model 8 predictions when using LR3PMS coded using Wellman and Estes terms. (B). Model 8 predictions when using LR3PMS coded using All Mental State terms.

Figure S21

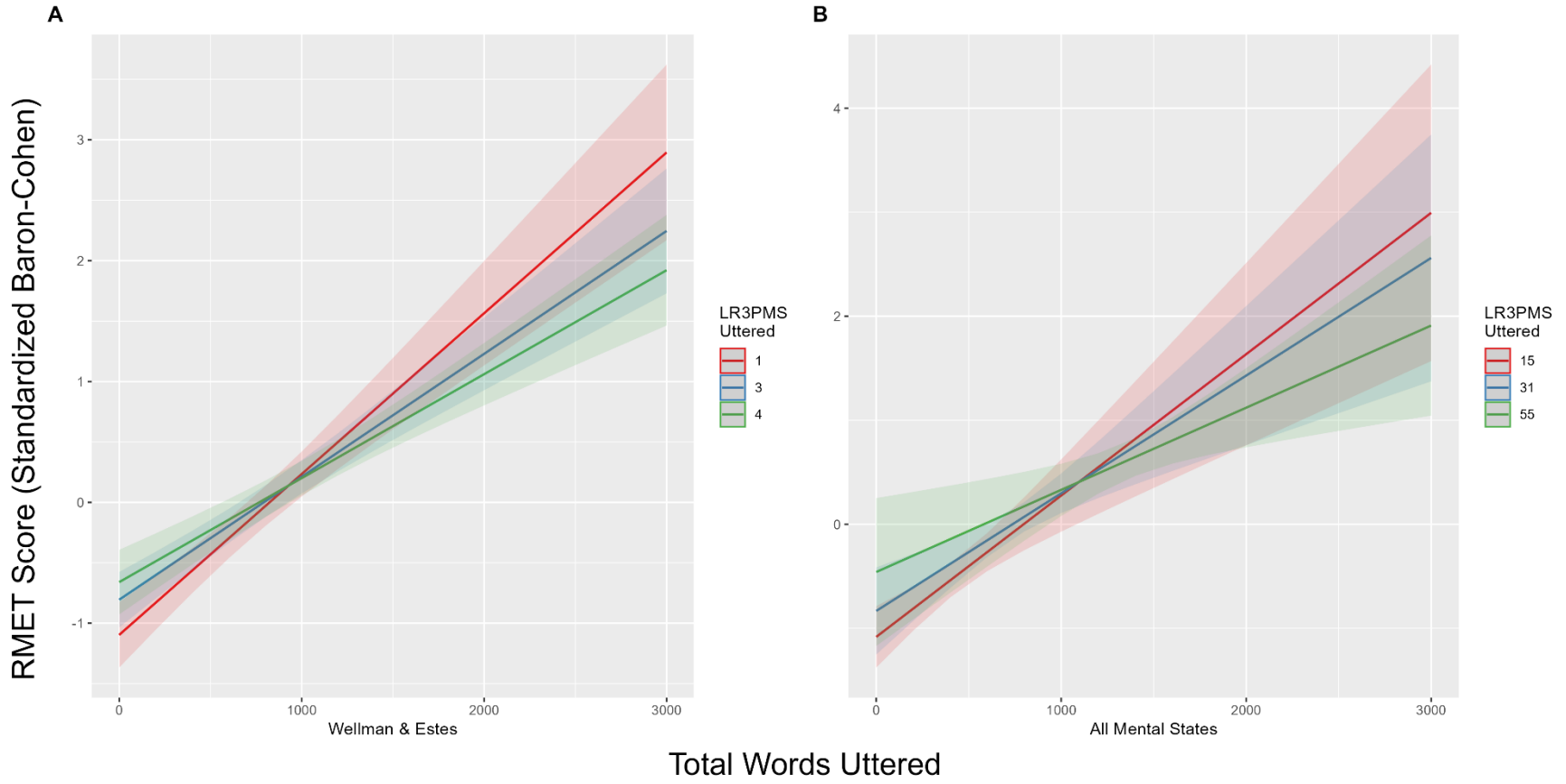
The Impact of Increased Counts of LR3PMS on RMET Scores is Attenuated as Total Words Uttered Increases



Note. (A) Predictions from Model 8 where LR3PMS were coded using Wellman and Estes terms. (B) Predictions from Model 8 where LR3PMS were coded using All Mental State terms. Values of Total Words Uttered corresponding to the lower quartile (395 words), the median (790 words), and the upper quartile (1261 words) were selected to examine the impact of increasing counts of LR3PMS uttered on RMET Score. Among the least talkative speakers, or those in the lower quartile of Total Words Uttered, as the count of LR3PMS uttered increased, performance on the RMET increased sharply (holding Total Words Uttered constant). A more modest, though still positive, effect was observed for participants who uttered the median value of Total Words Uttered. For the most talkative participants, or those in the upper quartile of Total Words Uttered, there was essentially no effect associated with a change in Welman and Estes terms LR3PMS (A) and a negative effect with a change in All Mental State terms LR3PMS (B).

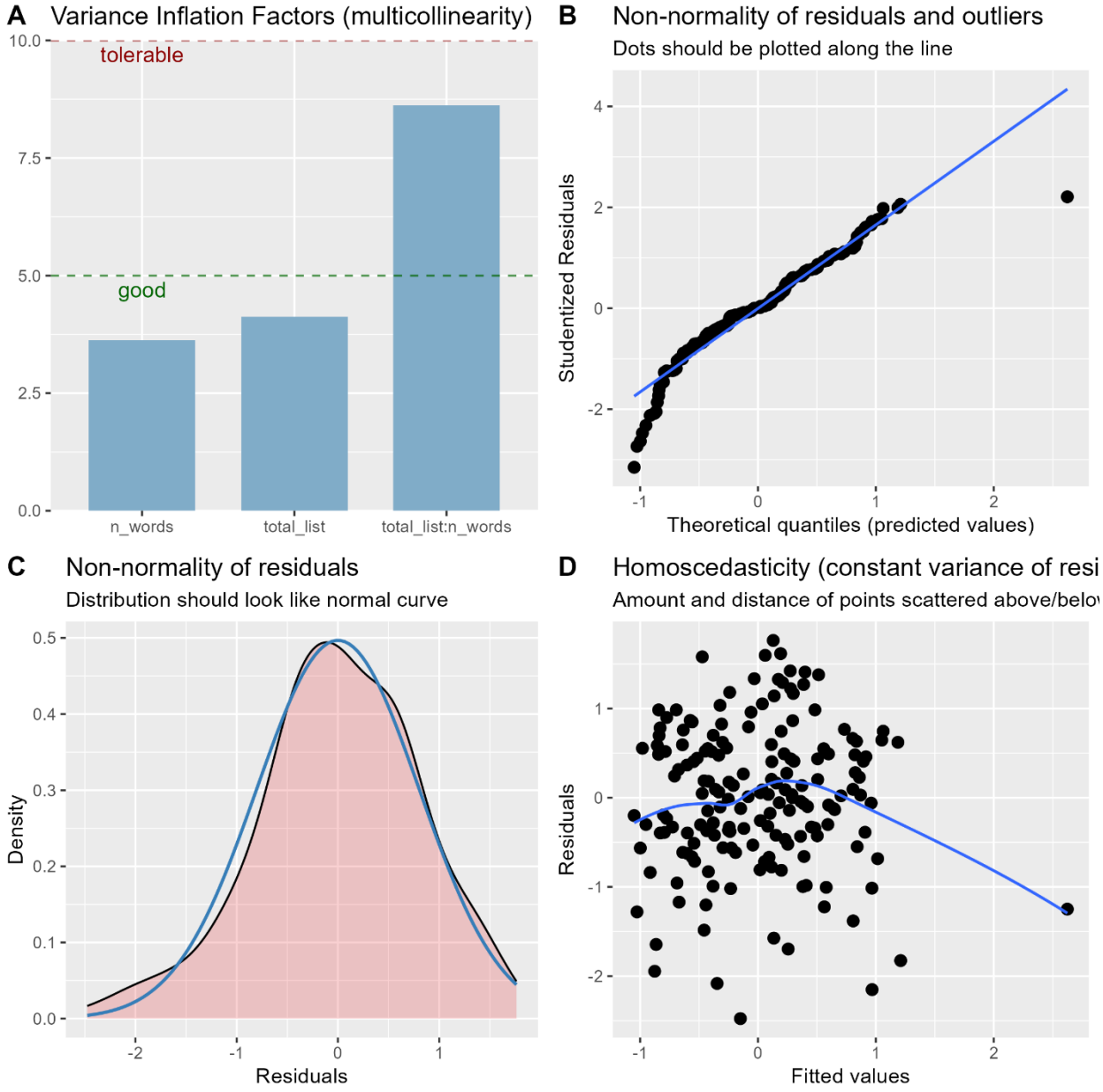
Figure S22

The Impact of Increased Counts of LR3PMS on RMET Scores is Attenuated as Total Words Uttered Increases



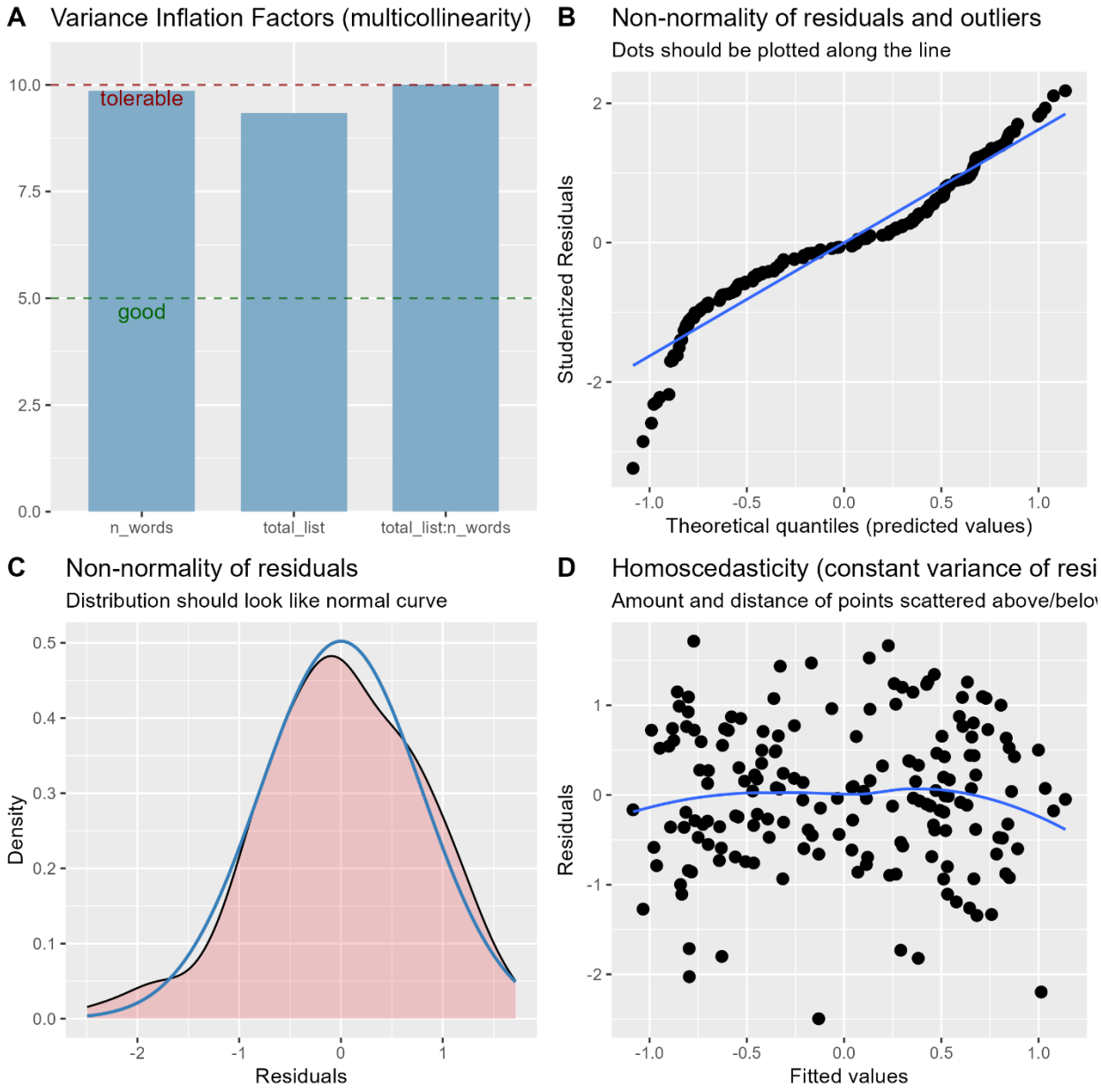
Note. *Note.* Values LR3PMS Uttered corresponding to the lower quartile (Wellman and Estes Terms = 1, All Mental State Terms = 15), the median (Wellman and Estes Terms = 3; All Mental State Terms = 31), and the upper quartile (Wellman and Estes Terms = 4; All Mental State Terms = 55) were selected to examine the impact of increasing counts of Total Words Uttered on RMET Score. Among participants who produced few LR3PMS (lower quartile), as the count of Total Words Uttered increased, performance on the RMET increased sharply (holding LR3PMS Uttered constant). A more modest, though still strongly positive, effect was observed for participants who uttered the median value of LR3PMS Uttered. For those participants who produced many LR3PMS (upper quartile), an even more modest though still fairly strongly positive effect on RMET score was observed. (A) LR3PMS coded with Wellman and Estes terms. (B) LR3PMS coded with All Mental State terms.

Figure S23



Note: Fit statistics for Model 8 on participant RMET performance using the Standardized Baron-Cohen coding scheme where LR3PMS were coded using the Wellman and Estes Terms coding scheme. (S23A) Variance inflation factors for each of the predictors in Model 8 show that they are uncorrelated. (S23B) QQ plot of model residuals shows that with the exception of some of the lower theoretical quantiles, residuals are normally distributed. (S23C) Another illustration of the residuals showing a normal distribution. (S23D) Variance in the residuals in more or less constant across the range of fitted values.

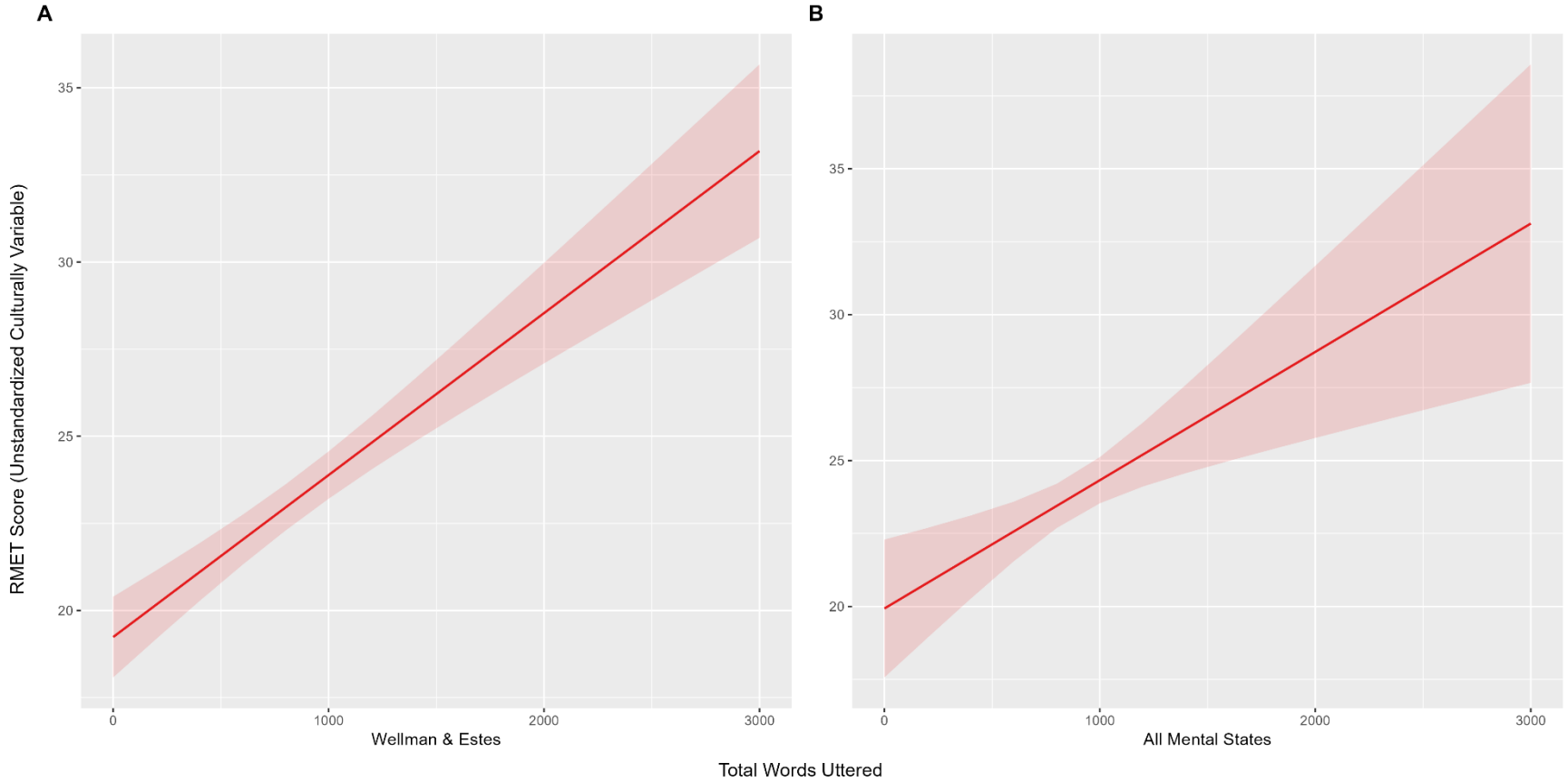
Figure S24



Note: Fit statistics for Model 8 on participant RMET performance using the Standardized Baron-Cohen coding scheme where LR3PMS were coded using the All Mental State Terms coding scheme. (S24A) Variance inflation factors for each of the predictors in Model 8 show that they are uncorrelated. (S24B) QQ plot of model residuals shows that with the exception of some of the lower theoretical quantiles, residuals are normally distributed. (S24C) Another illustration of the residuals showing a normal distribution. (S24D) Variance in the residuals in more or less constant across the range of fitted values.

Figure S25

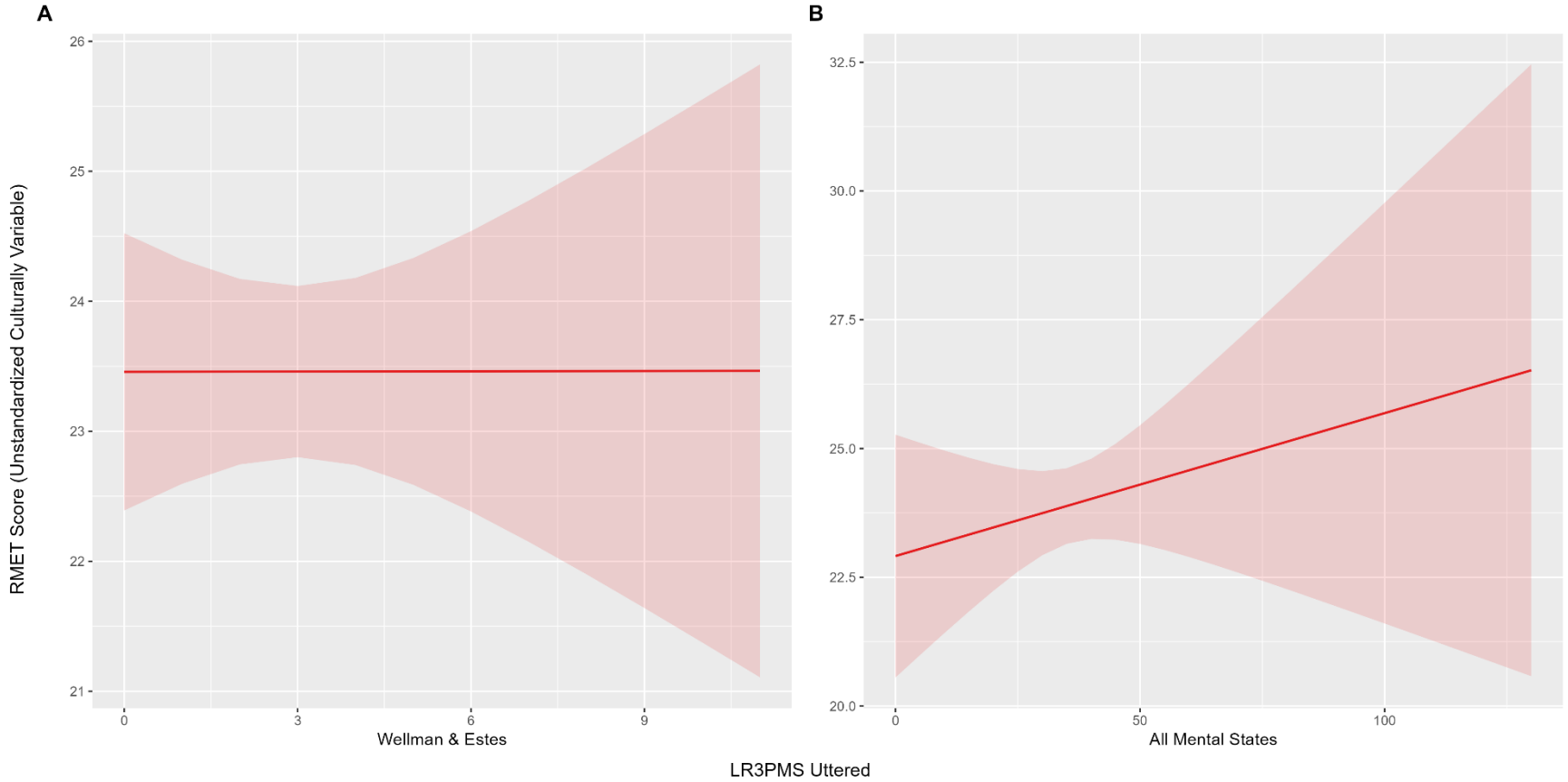
Total Words Uttered by Participants Strongly Positively Predicts Performance on RMET



Note. The predicted results of Model 8 suggest that as the counts of Total Words Uttered increase, so too do participant scores on the Reading the Mind in the Eyes Test (RMET) using the Unstandardized Culturally Variable coding scheme. Visualization holds LR3PMS Uttered constant at the sample mean value. (A) Model 8 predictions when using LR3PMS coded using Wellman and Estes terms. (B). Model 8 predictions when using LR3PMS coded using All Mental State terms.

Figure S26

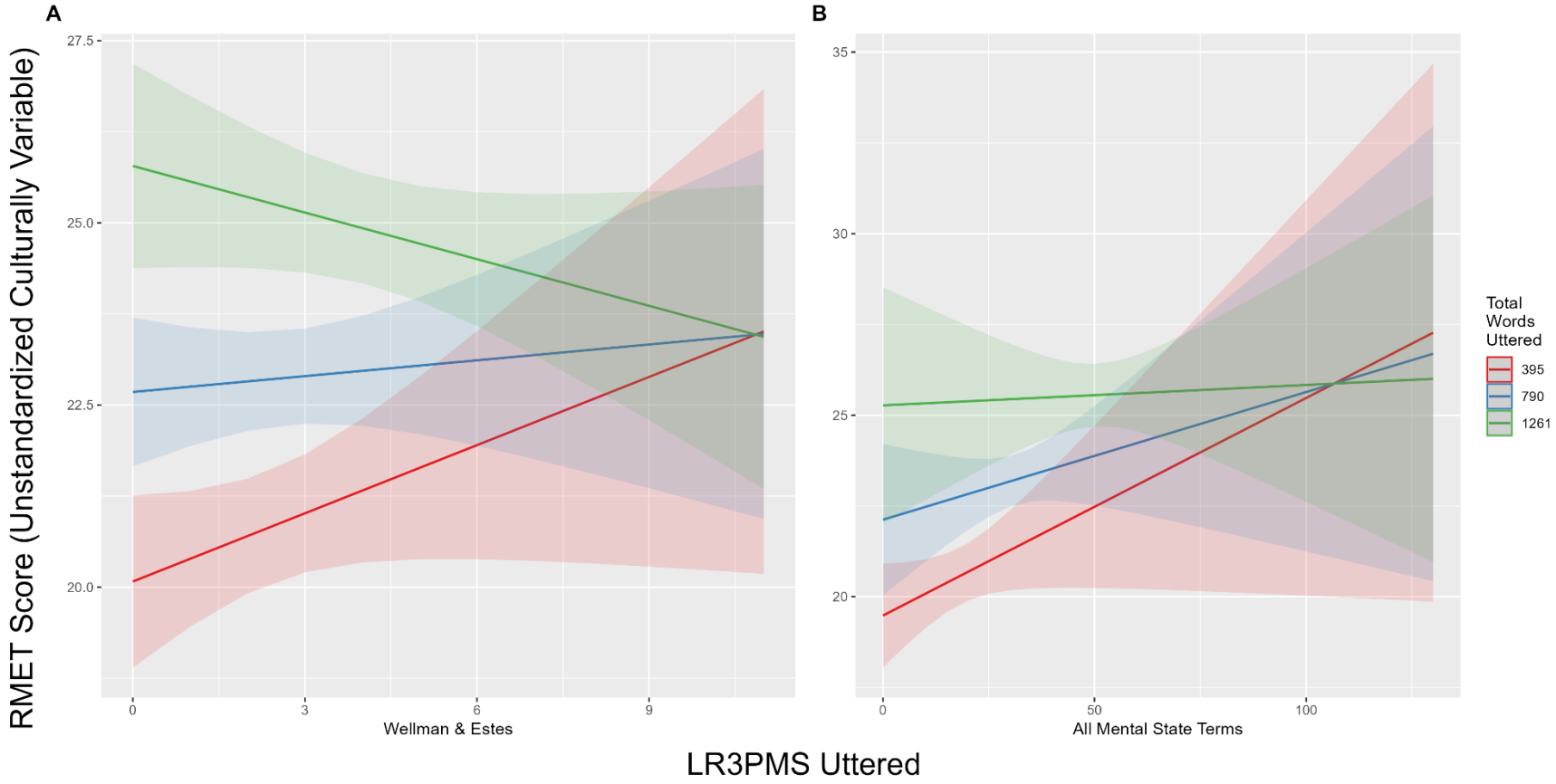
LR3PMS Uttered by Participants Predicts Performance on RMET Less Strongly Than Total Words Uttered



Note. Model 8 predicts that the participants' scores on the RMET will increase modestly as the total number of All Mental State Terms LR3PMS uttered increases. This effect is independent of, albeit weaker than that of Total Words Uttered. Visualizations hold Total Words Uttered constant at the sample mean value. (A) Model 8 predictions when using LR3PMS coded using Wellman and Estes terms. (B). Model 8 predictions when using LR3PMS coded using All Mental State terms.

Figure S27

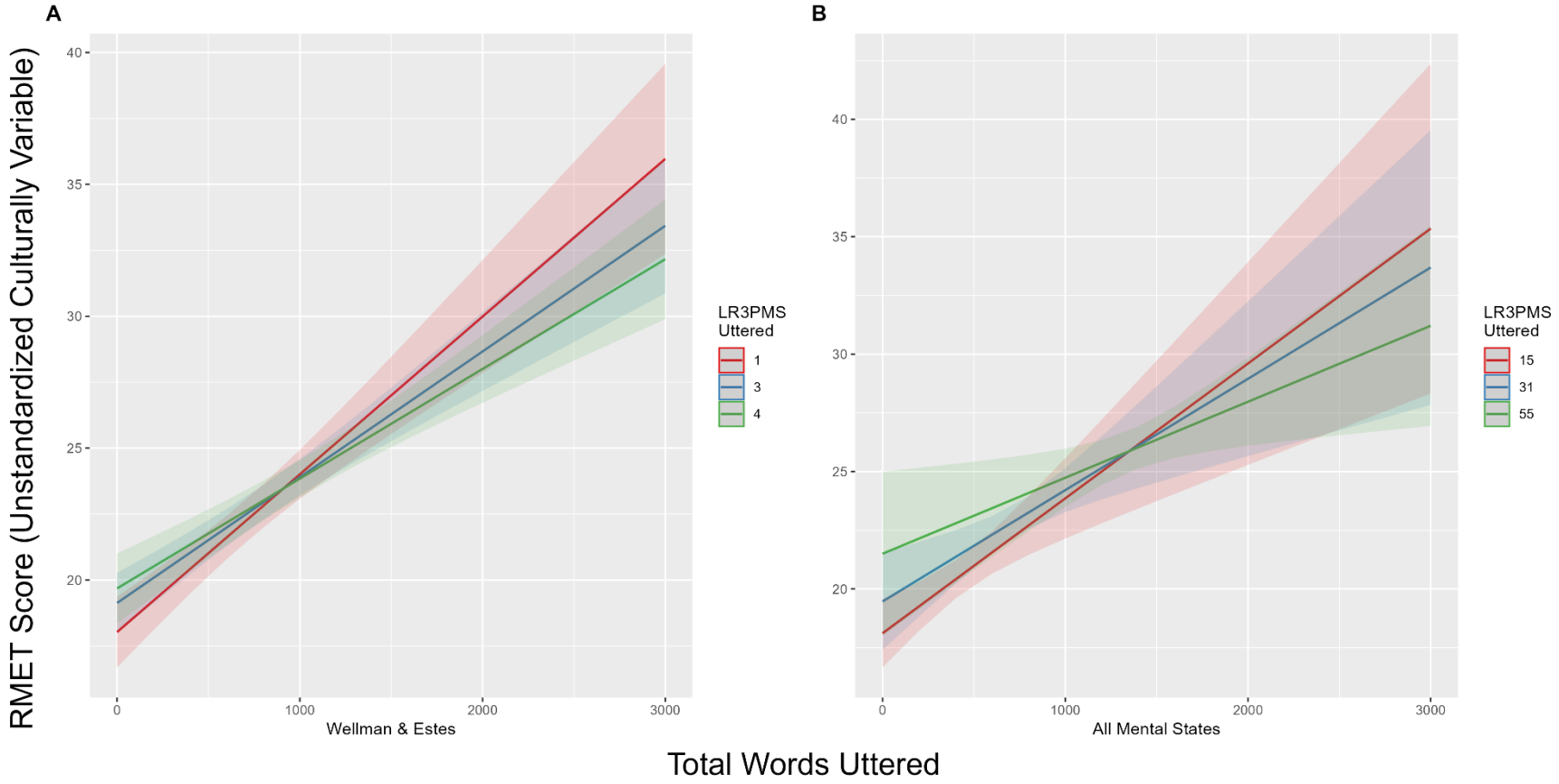
The Impact of Increased Counts of LR3PMS on RMET Scores is Attenuated as Total Words Uttered Increases



Note. (A) Predictions from Model 8 where LR3PMS were coded using Wellman and Estes terms. (B) Predictions from Model 8 where LR3PMS were coded using All Mental State terms. Values of Total Words Uttered corresponding to the lower quartile (395 words), the median (790 words), and the upper quartile (1261 words) were selected to examine the impact of increasing counts of LR3PMS uttered on RMET Score. Among the least talkative speakers, or those in the lower quartile of Total Words Uttered, as the count of LR3PMS uttered increased, performance on the RMET increased sharply (holding Total Words Uttered constant). A more modest, though still positive, effect was observed for participants who uttered the median value of Total Words Uttered. For the most talkative participants, or those in the upper quartile of Total Words Uttered, there was essentially no effect associated with a change in Welman and Estes terms LR3PMS (A) and a negative effect with a change in All Mental State terms LR3PMS (B).

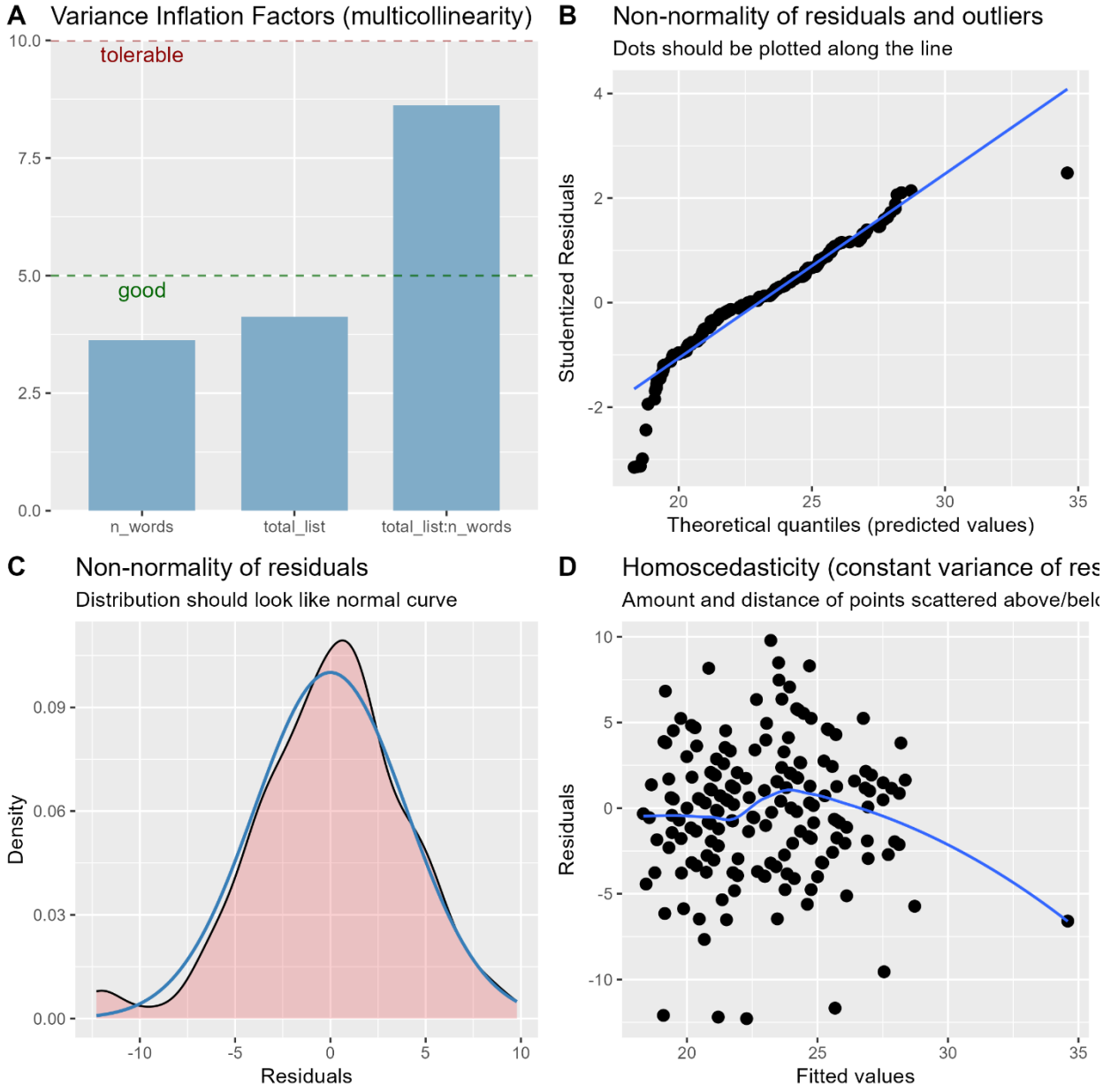
Figure S28

The Impact of Increased Counts of LR3PMS on RMET Scores is Attenuated as Total Words Uttered Increases



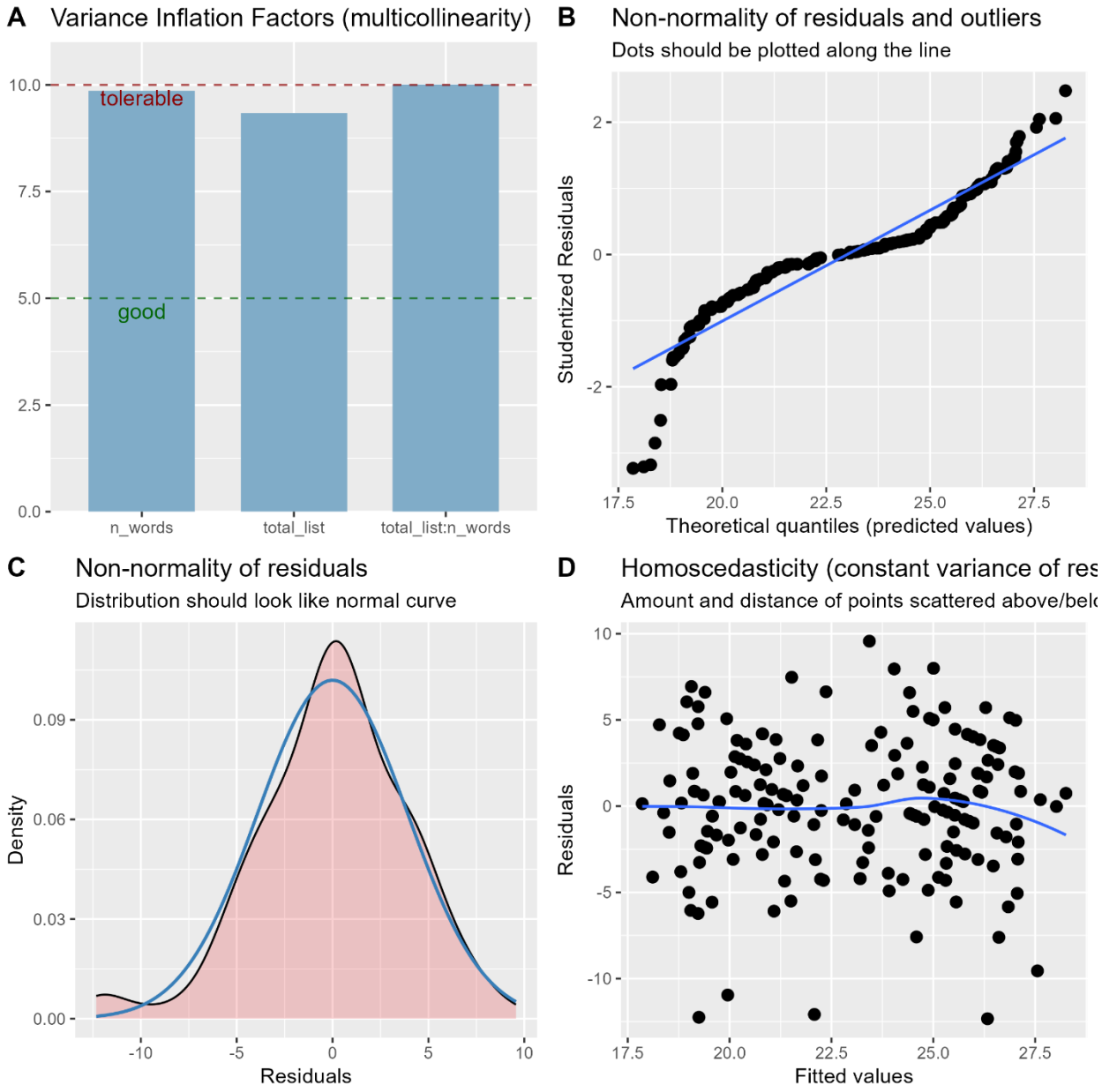
Note. *Note.* Values LR3PMS Uttered corresponding to the lower quartile (Wellman and Estes Terms = 1, All Mental State Terms = 15), the median (Wellman and Estes Terms = 3; All Mental State Terms = 31), and the upper quartile (Wellman and Estes Terms = 4; All Mental State Terms = 55) were selected to examine the impact of increasing counts of Total Words Uttered on RMET Score. Among participants who produced few LR3PMS (lower quartile), as the count of Total Words Uttered increased, performance on the RMET increased sharply (holding LR3PMS Uttered constant). A more modest, though still strongly positive, effect was observed for participants who uttered the median value of LR3PMS Uttered. For those participants who produced many LR3PMS (upper quartile), an even more modest though still fairly strongly positive effect on RMET score was observed. (A) LR3PMS coded with Wellman and Estes terms. (B) LR3PMS coded with All Mental State terms.

Figure S29



Note: Fit statistics for Model 8 on participant RMET performance using the Unstandardized Culturally Variable coding scheme where LR3PMS were coded using the Wellman and Estes Terms coding scheme. (S29A) Variance inflation factors for each of the predictors in Model 8 show that they are uncorrelated. (S29B) QQ plot of model residuals shows that with the exception of some of the lower theoretical quantiles, residuals are normally distributed. (S29C) Another illustration of the residuals showing a normal distribution. (S29D) Variance in the residuals in more or less constant across the range of fitted values.

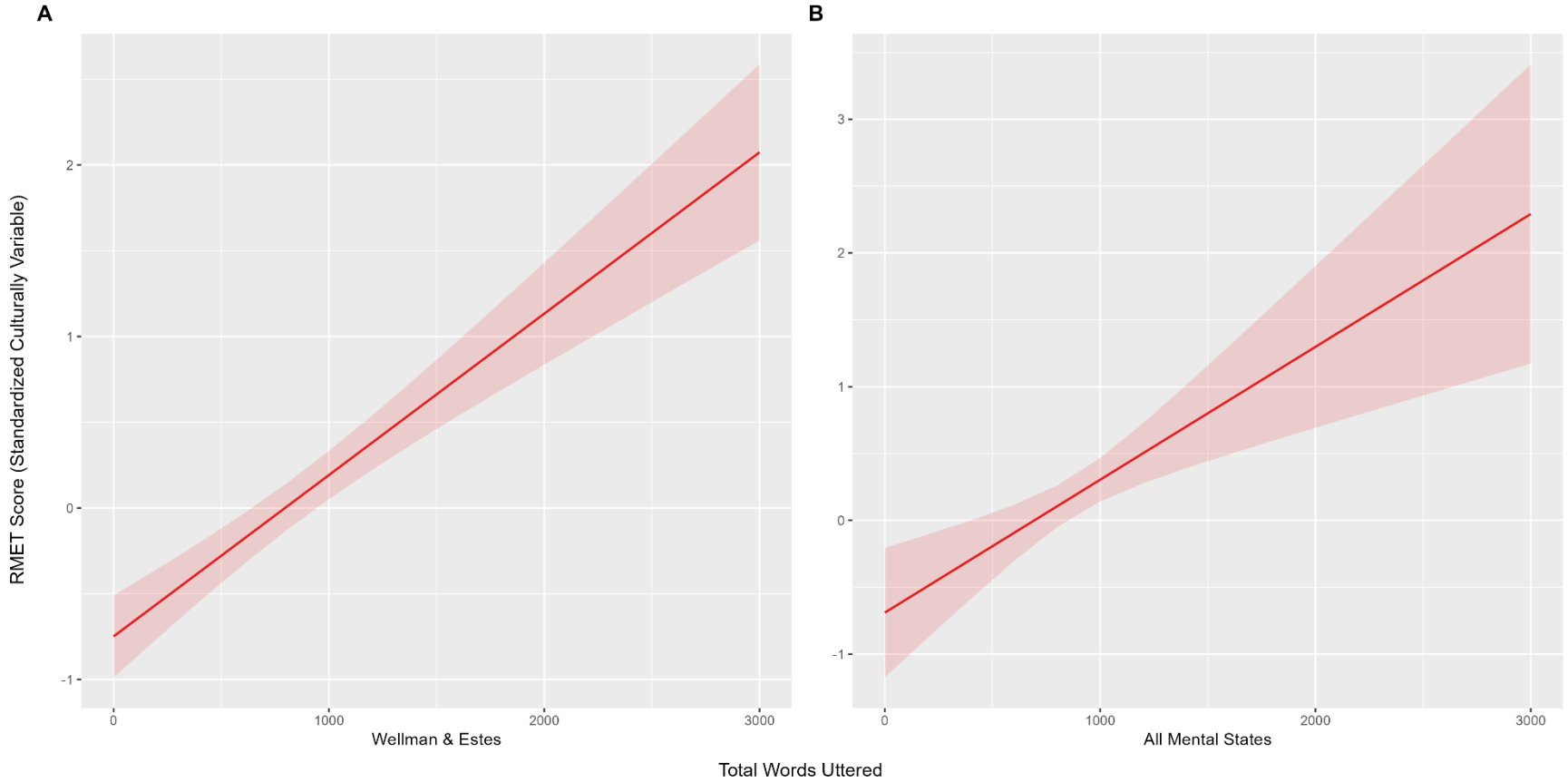
Figure S30



Note: Fit statistics for Model 8 on participant RMET performance using the Unstandardized Culturally Variable coding scheme where LR3PMS were coded using the All Mental State Terms coding scheme. (S30A) Variance inflation factors for each of the predictors in Model 8 show that they are uncorrelated. (S30B) QQ plot of model residuals shows that with the exception of some of the lower theoretical quantiles, residuals are normally distributed. (S30C) Another illustration of the residuals showing a normal distribution. (S30D) Variance in the residuals in more or less constant across the range of fitted values.

Figure S31

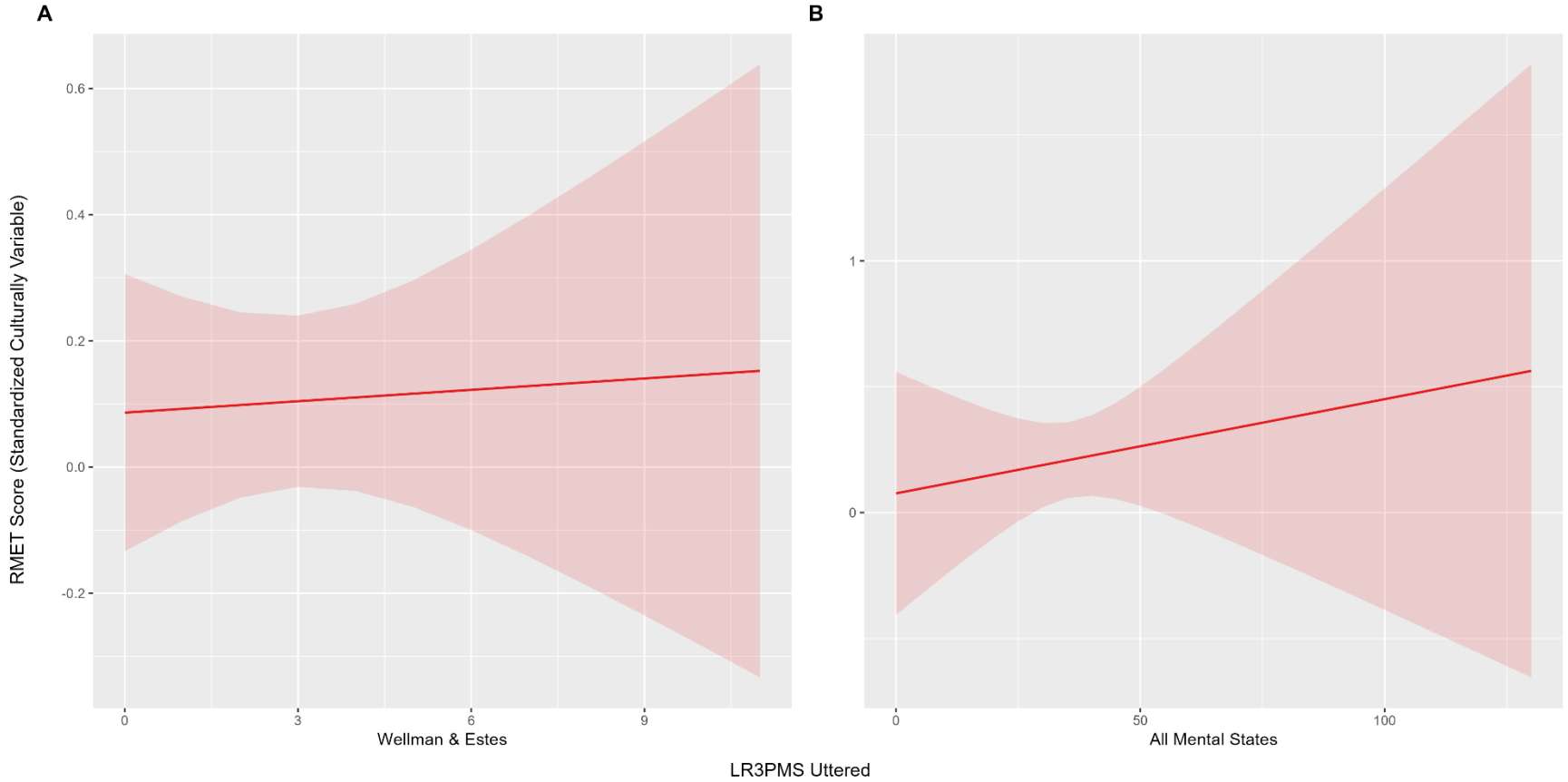
Total Words Uttered by Participants Strongly Positively Predicts Performance on RMET



Note. The predicted results of Model 8 suggest that as the counts of Total Words Uttered increase, so too do participant scores on the Reading the Mind in the Eyes Test (RMET) using the Standardized Culturally Variable coding scheme. Visualization holds LR3PMS Uttered constant at the sample mean value. (A) Model 8 predictions when using LR3PMS coded using Wellman and Estes terms. (B). Model 8 predictions when using LR3PMS coded using All Mental State terms.

Figure S32

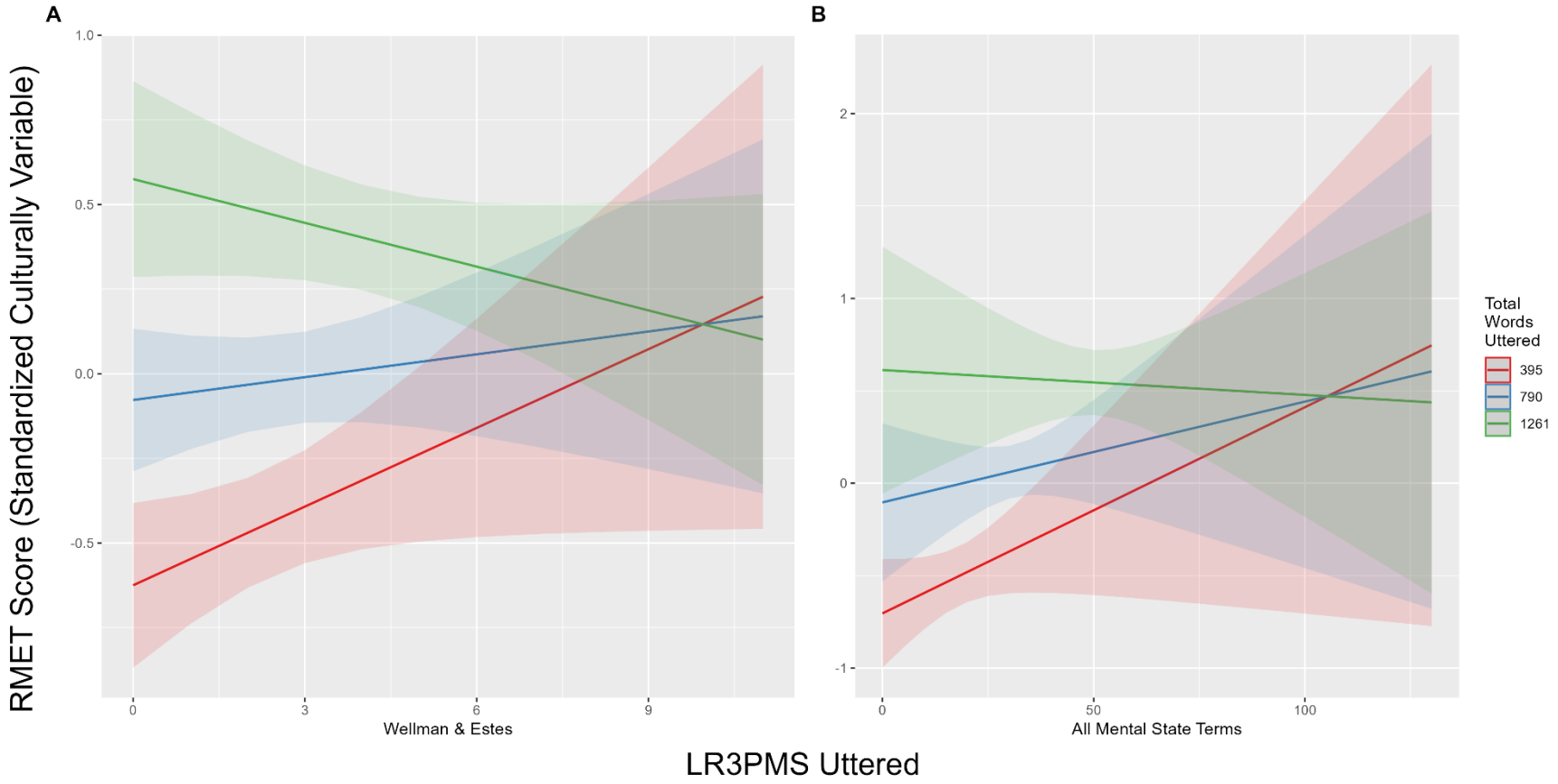
LR3PMS Uttered by Participants Predicts Performance on RMET Less Strongly Than Total Words Uttered



Note. Model 8 predicts that the participants' scores on the RMET will increase modestly as the total number of All Mental State Terms LR3PMS uttered increases. This effect is independent of, albeit weaker than that of Total Words Uttered. Visualizations hold Total Words Uttered constant at the sample mean value. (A) Model 8 predictions when using LR3PMS coded using Wellman and Estes terms. (B). Model 8 predictions when using LR3PMS coded using All Mental State terms.

Figure S33

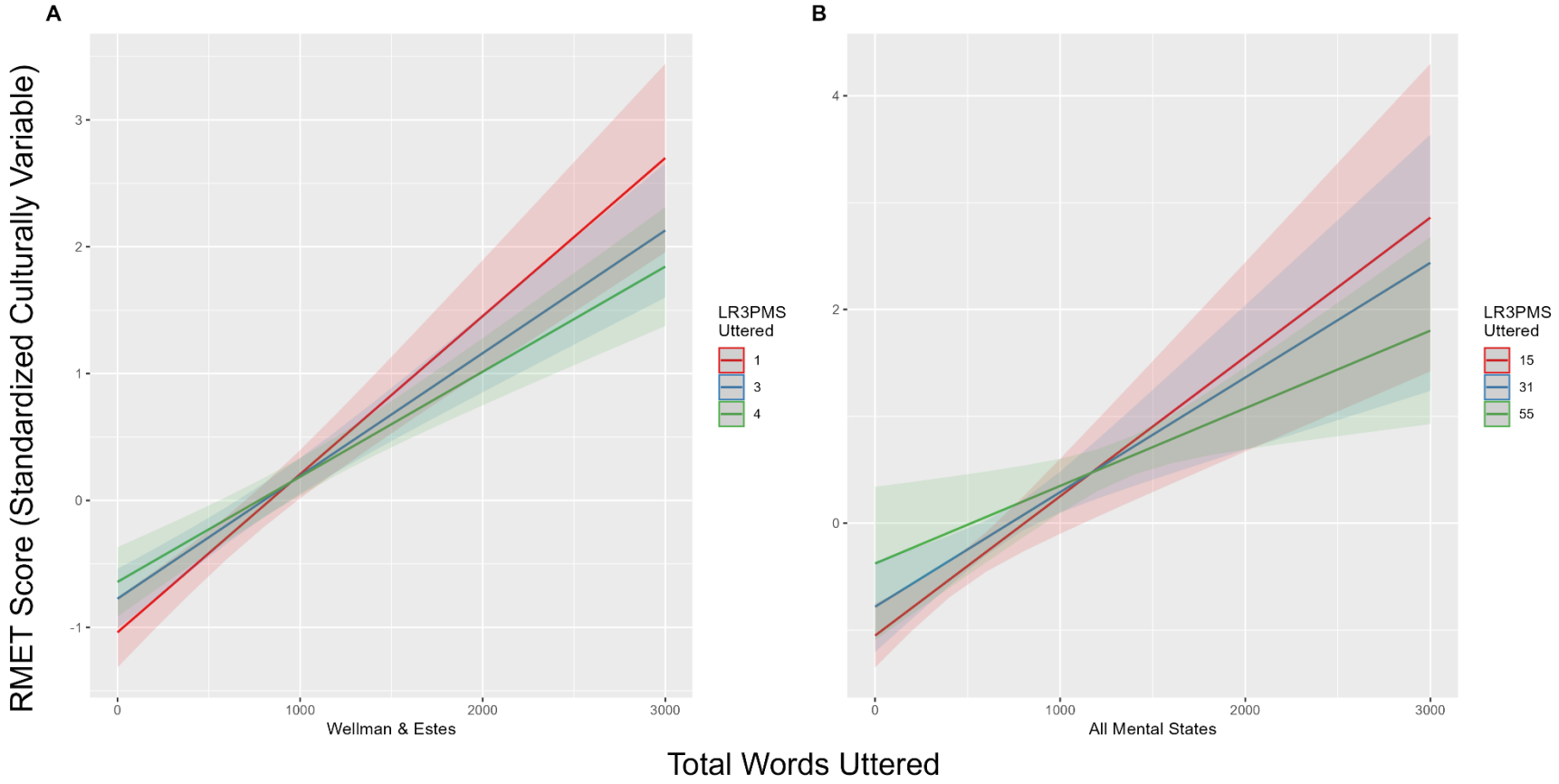
The Impact of Increased Counts of LR3PMS on RMET Scores is Attenuated as Total Words Uttered Increases



Note. (A) Predictions from Model 8 where LR3PMS were coded using Wellman and Estes terms. (B) Predictions from Model 8 where LR3PMS were coded using All Mental State terms. Values of Total Words Uttered corresponding to the lower quartile (395 words), the median (790 words), and the upper quartile (1261 words) were selected to examine the impact of increasing counts of LR3PMS uttered on RMET Score. Among the least talkative speakers, or those in the lower quartile of Total Words Uttered, as the count of LR3PMS uttered increased, performance on the RMET increased sharply (holding Total Words Uttered constant). A more modest, though still positive, effect was observed for participants who uttered the median value of Total Words Uttered. For the most talkative participants, or those in the upper quartile of Total Words Uttered, there was essentially no effect associated with a change in Welman and Estes terms LR3PMS (A) and a negative effect with a change in All Mental State terms LR3PMS (B).

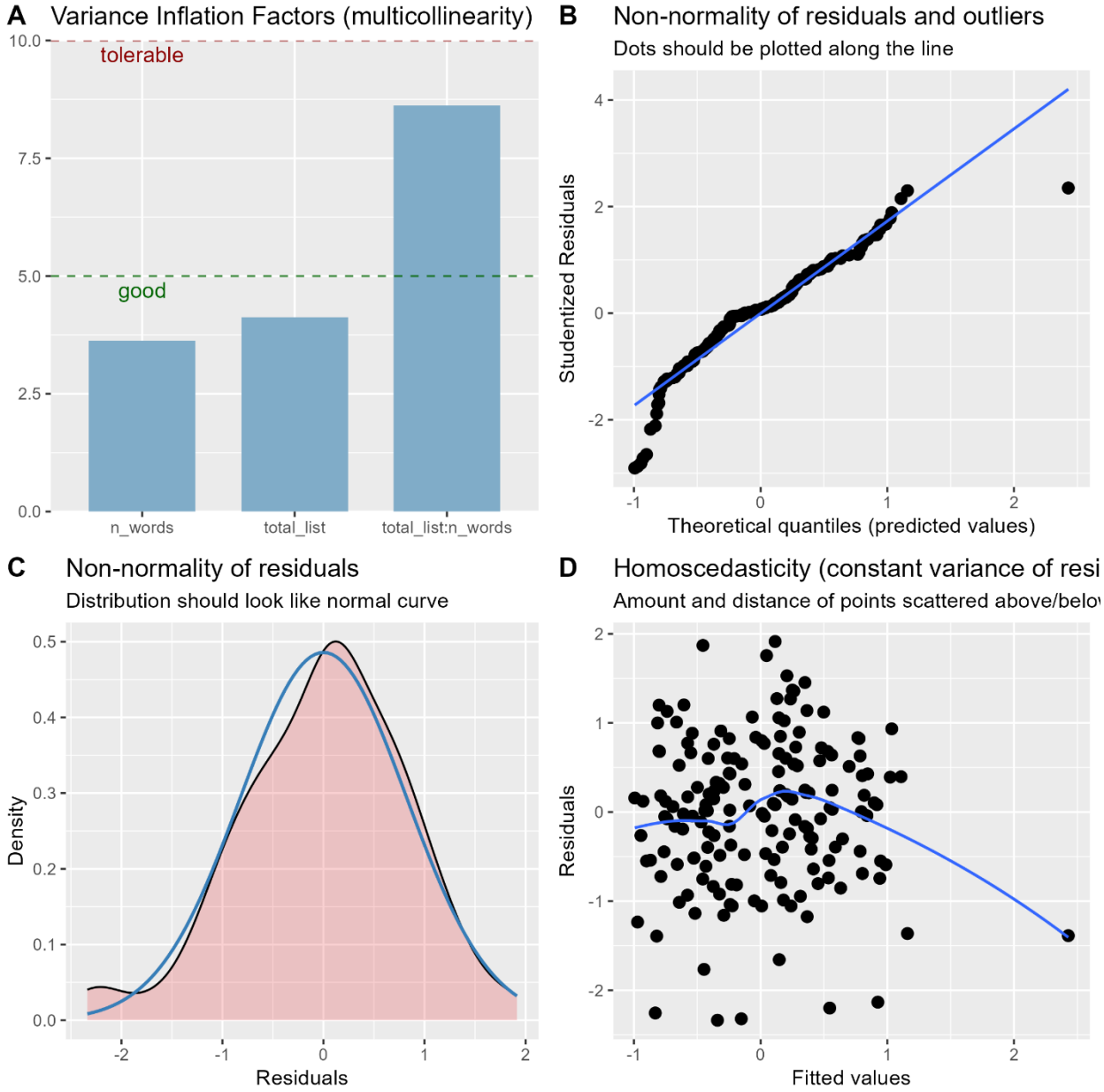
Figure S34

The Impact of Increased Counts of LR3PMS on RMET Scores is Attenuated as Total Words Uttered Increases



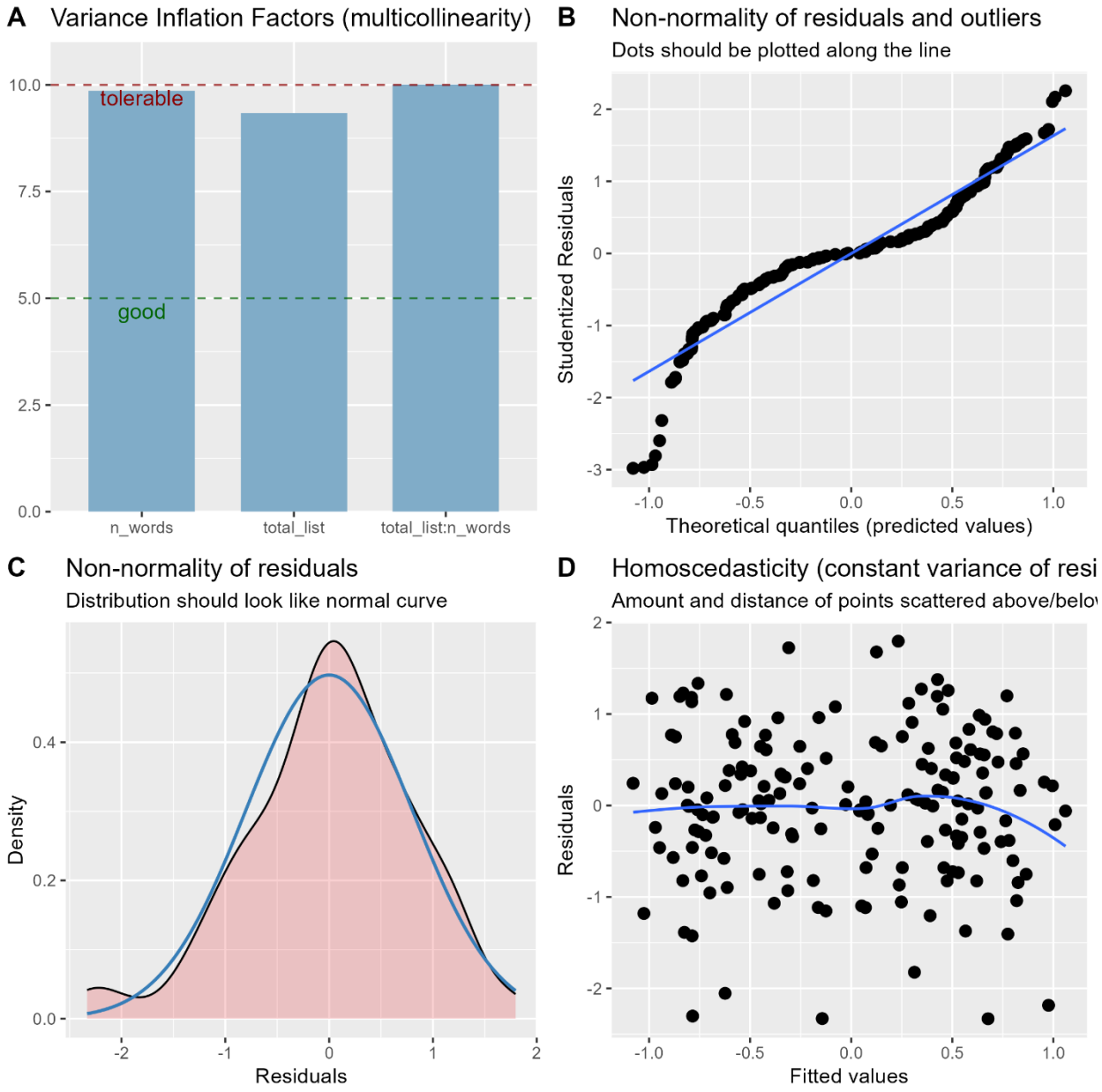
Note. *Note.* Values LR3PMS Uttered corresponding to the lower quartile (Wellman and Estes Terms = 1, All Mental State Terms = 15), the median (Wellman and Estes Terms = 3; All Mental State Terms = 31), and the upper quartile (Wellman and Estes Terms = 4; All Mental State Terms = 55) were selected to examine the impact of increasing counts of Total Words Uttered on RMET Score. Among participants who produced few LR3PMS (lower quartile), as the count of Total Words Uttered increased, performance on the RMET increased sharply (holding LR3PMS Uttered constant). A more modest, though still strongly positive, effect was observed for participants who uttered the median value of LR3PMS Uttered. For those participants who produced many LR3PMS (upper quartile), an even more modest though still fairly strongly positive effect on RMET score was observed. (A) LR3PMS coded with Wellman and Estes terms. (B) LR3PMS coded with All Mental State terms.

Figure S35



Note: Fit statistics for Model 8 on participant RMET performance using the Standardized Culturally Variable coding scheme where LR3PMS were coded using the Wellman and Estes Terms coding scheme. (S35A) Variance inflation factors for each of the predictors in Model 8 show that they are uncorrelated. (S35B) QQ plot of model residuals shows that with the exception of some of the lower theoretical quantiles, residuals are normally distributed. (S35C) Another illustration of the residuals showing a normal distribution. (S35D) Variance in the residuals in more or less constant across the range of fitted values.

Figure S36



Note: Fit statistics for Model 8 on participant RMET performance using the Standardized Culturally Variable coding scheme where LR3PMS were coded using the All Mental State Terms coding scheme. (S36A) Variance inflation factors for each of the predictors in Model 8 show that they are uncorrelated. (S36B) QQ plot of model residuals shows that with the exception of some of the lower theoretical quantiles, residuals are normally distributed. (S36C) Another illustration of the residuals showing a normal distribution. (S36D) Variance in the residuals in more or less constant across the range of fitted values.

References

Adams Jr, R. B., Rule, N. O., Franklin Jr, R. G., Wang, E., Stevenson, M. T., Yoshikawa, S., Nomura, M., Sato, W., Kveraga, K., & Ambady, N. (2010). Cross-cultural reading the mind in the eyes: An fMRI investigation. *Journal of Cognitive Neuroscience*, *22*(1), 97–108.

Aikhenvald, A. Y. (2004). *Evidentiality*. Oxford University Press.

Aikhenvald, A. Y., & Dixon, R. M. W. (2003). *Studies in evidentiality* (Vol. 54). John Benjamins Publishing.

Aival-Naveh, E., Rothschild-Yakar, L., & Kurman, J. (2019). Keeping culture in mind: A systematic review and initial conceptualization of mentalizing from a cross-cultural perspective. *Clinical Psychology: Science and Practice*, *26*(4), e12300.

Akhtar, N., & Gernsbacher, M. A. (2007). Joint Attention and Vocabulary Development: A Critical Look. *Language and Linguistics Compass*, *1*(3), 195–207. https://doi.org/10.1111/j.1749-818X.2007.00014.x

Apperly, I. A. (2008). Beyond Simulation–Theory and Theory–Theory: Why social cognitive neuroscience should use its own concepts to study theory of mind. *Cognition*, *107*(1), 266–283.

Apperly, I. A. (2012). What is “theory of mind”? Concepts, cognitive processes and individual differences. *Quarterly Journal of Experimental Psychology*, *65*(5), 825–839. https://doi.org/10.1080/17470218.2012.676055

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, *116*(4), 953.

Arab Barometer. (2019). *Arab Barometer V - Morocco Country Report*. Arab Barometer. https://www.arabbarometer.org/wp-content/uploads/ABV_Morocco_Report_Public-Opinion_Arab-Barometer_2019.pdf

Avis, J., & Harris, P. L. (1991). Belief-desire reasoning among Baka children: Evidence for a universal conception of mind. *Child Development*, *62*(3), 460–467.

Baetens, K., Ma, N., Steen, J., & Van Overwalle, F. (2014). Involvement of the mentalizing network in social and non-social high construal. *Social Cognitive and Affective Neuroscience*, *9*(6), 817–824.

Baez, S., Tangarife, M. A., Davila-Mejia, G., Trujillo-Güiza, M., & Forero, D. A. (2023). Performance in emotion recognition and theory of mind tasks in social anxiety and generalized anxiety disorders: A systematic review and meta-analysis. *Frontiers in Psychiatry*, *14*, 1192683.

Bagg, E., Pickard, H., Tan, M., Smith, T. J., Simonoff, E., Pickles, A., Carter Leno, V., & Bedford, R. (2024). Testing the social motivation theory of autism: The role of co-occurring anxiety. *Journal of Child Psychology and Psychiatry*, *65*(7), 899–909.

Balcetis, E. (2016). Approach and avoidance as organizing structures for motivated distance perception. *Emotion Review*, *8*(2), 115–128.

Baron-Cohen, S. (1997a). How to build a baby that can read minds: Cognitive mechanisms in mindreading. *The Maladapted Mind: Classic Readings in Evolutionary Psychopathology*, 207–239.

Baron-Cohen, S. (1997b). *Mindblindness: An essay on autism and theory of mind*. MIT press.

Baron-Cohen, S., Jolliffe, T., Mortimore, C., & Robertson, M. (1997). Another advanced test of theory of mind: Evidence from very high functioning adults with autism or Asperger syndrome. *Journal of Child Psychology and Psychiatry*, *38*(7), 813–822.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a theory of mind? *Cognition*, *21*(1), 37–46.

Baron-Cohen, S., O'Riordan, M., Jones, R., Stone, V., & Plaisted, K. (1999). A new test of social sensitivity: Detection of faux pas in normal children and children with Asperger syndrome. *Journal of Autism and Developmental Disorders*, *29*(5), 407–418.

Baron-Cohen, S., Richler, J., Bisarya, D., Gurunathan, N., & Wheelwright, S. (2003). The systemizing quotient: An investigation of adults with Asperger syndrome or high–functioning autism, and normal sex differences. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*(1430), 361–374.

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

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *42*(2), 241–251.

Barrett, H. C. (2005). Adaptations to predators and prey. *The Handbook of Evolutionary Psychology*, 200–223.

Barrett, H. C. (2015). *The shape of thought: How mental adaptations evolve*. Oxford University Press.

Barrett, H. C., Bolyanatz, A., Crittenden, A. N., Fessler, D. M., Fitzpatrick, S., Gurven, M., Henrich, J., Kanovsky, M., Kushnick, G., Pisor, A., & others. (2016). Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of Sciences*, *113*(17), 4688–4693.

Barrett, H. C., Broesch, T., Scott, R. M., He, Z., Baillargeon, R., Wu, D., Bolz, M., Henrich, J., Setoh, P., Wang, J., & others. (2013). Early false-belief understanding in traditional non-

Western societies. *Proceedings of the Royal Society of London B: Biological Sciences*, *280*(1755), 20122654.

Barrett, H. C., Cosmides, L., & Tooby, J. (2010). Coevolution of cooperation, causal cognition and mindreading. *Communicative & Integrative Biology*, *3*(6), 522–524.

Barron, A., & Schneider, K. P. (2009). *Variational pragmatics: Studying the impact of social factors on language use in interaction*.

Bendix, E. H. (1992). The grammaticalization of responsibility and evidence: Interactional manipulation of evidential categories in Newari. In J. H. Hill & J. T. Irvine (Eds.), *Responsibility and evidence in oral discourse* (pp. 226–247).

Benson-Amram, S., Griebling, H. J., & Sluka, C. M. (2023). The current state of carnivore cognition. *Animal Cognition*, *26*(1), 37–58.

Berge, M. T., & Raad, B. D. (2001). The construction of a joint taxonomy of traits and situations. *European Journal of Personality*, *15*(4), 253–276.

Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. Univ of California Press.

Berriane, M., de Haas, H., & Natter, K. (2021). *Social transformations and migrations in Morocco*. International Migration Institute network (IMI).

Bickerton, D. (1984). The language bioprogram hypothesis. *Behavioral and Brain Sciences*, *7*(2), 173–188.

Bjornsdottir, R. T., & Rule, N. O. (2016). On the relationship between acculturation and intercultural understanding: Insight from the Reading the Mind in the Eyes test. *International Journal of Intercultural Relations*, *52*, 39–48. https://doi.org/10.1016/j.ijintrel.2016.03.003

Black, J. E. (2019). An IRT analysis of the Reading the Mind in the Eyes test. *Journal of Personality Assessment*, *101*(4), 425–433.

Bloom, L., Rispoli, M., Gartner, B., & Hafitz, J. (1989). Acquisition of complementation. *Journal of Child Language*, *16*(1), 101–120.

Bloom, P., & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, *77*(1), B25–B31.

Boas, F. (1911). *Handbook of American Indian Languages* (Bulletin 40, Vol. 1). Government Print Office (Smithsonian Institution, Bureau of American Ethnology).

Booth, J. R., Hall, W. S., Robison, G. C., & Kim, S. Y. (1997). Acquisition of the mental state verb know by 2-to 5-year-old children. *Journal of Psycholinguistic Research*, *26*(6), 581–603.

Borghi, A. M., Binkofski, F., Castelfranchi, C., Cimatti, F., Scorolli, C., & Tummolini, L. (2017). The challenge of abstract concepts. *Psychological Bulletin*, *143*(3), 263.

Bornstein, M. H. (2006). Hue categorization and color naming: Physics to sensation to perception. *Progress in Colour Studies: Volume II. Psychological Aspects*, 35.

Bornstein, M. H., Kessen, W., & Weiskopf, S. (1976). The categories of hue in infancy. *Science*, *191*(4223), 201–202.

Boroditsky, L. (2011). How language shapes thought. *Scientific American*, *304*(2), 62–65.

Boyd, R., & Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology*, *13*(3), 171–195.

Bradford, E. E., Jentzsch, I., Gomez, J.-C., Chen, Y., Zhang, D., & Su, Y. (2018). Cross-cultural differences in adult Theory of Mind abilities: A comparison of native-English speakers and native-Chinese speakers on the Self/Other Differentiation task. *Quarterly Journal of Experimental Psychology*, 1747021818757170.

Bräuer, J., Call, J., & Tomasello, M. (2005). All great ape species follow gaze to distant locations and around barriers. *Journal of Comparative Psychology*, *119*(2), 145.

Bretherton, I., & Beeghly, M. (1982). Talking about internal states: The acquisition of an explicit theory of mind. *Developmental Psychology*, *18*(6), 906.

Brooks, R., & Meltzoff, A. N. (2005). The development of gaze following and its relation to language. *Developmental Science*, *8*(6), 535–543.

Brooks, R., & Meltzoff, A. N. (2015). Connecting the dots from infancy to childhood: A longitudinal study connecting gaze following, language, and explicit theory of mind. *Journal of Experimental Child Psychology*, *130*, 67–78.

Brown, J. R., Donelan-McCall, N., & Dunn, J. (1996). Why talk about mental states? The significance of children's conversations with friends, siblings, and mothers. *Child Development*, *67*(3), 836–849.

Brown, R. W., & Lenneberg, E. H. (1954). A study in language and cognition. *The Journal of Abnormal and Social Psychology*, *49*(3), 454–462. https://doi.org/10.1037/h0057814

Bugnyar, T., Reber, S. A., & Buckner, C. (2016). Ravens attribute visual access to unseen competitors. *Nature Communications*, *7*, 10506.

Burgess, N. (2006). Spatial memory: How egocentric and allocentric combine. *Trends in Cognitive Sciences*, *10*(12), 551–557.

Buss, D. M., Larsen, R. J., Westen, D., & Semmelroth, J. (1992). Sex differences in jealousy: Evolution, physiology, and psychology. *Psychological Science*, *3*(4), 251–256.

Buss, D. M., Shackelford, T. K., Kirkpatrick, L. A., Choe, J. C., Lim, H. K., Hasegawa, M., Hasegawa, T., & Bennett, K. (1999). Jealousy and the nature of beliefs about infidelity: Tests of competing hypotheses about sex differences in the United States, Korea, and Japan. *Personal Relationships*, *6*(1), 125–150.

Butterfill, S. A., & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind & Language*, *28*(5), 606–637.

Buunk, B. P., Angleitner, A., Oubaid, V., & Buss, D. M. (1996). Sex differences in jealousy in evolutionary and cultural perspective: Tests from the Netherlands, Germany, and the United States. *Psychological Science*, *7*(6), 359–363.

Caballero, F. S., i Sellabona, E. S., Serrano, J., Sánchez, C. R., Caño, A., & Codony, A. A. (2013). Let's share perspectives! Mentalistic skills involved in cooperation. *International Journal of Educational Psychology: IJEP*, *2*(3), 325–352.

Cain, W. S., Stevens, J. C., Nickou, C. M., Giles, A., Johnston, I., & Garcia-Medina, M. R. (1994). Life-span development of odor identification, learning, and olfactory sensitivity. *Perception*, *24*(12), 1457–1472.

Call, J., & Tomasello, M. (1998). Distinguishing intentional from accidental actions in orangutans (Pongo pygmaeus), chimpanzees (Pan troglodytes) and human children (Homo sapiens). *Journal of Comparative Psychology*, *112*(2), 192.

Callaghan, T., Rochat, P., Lillard, A., Claux, M. L., Odden, H., Itakura, S., Tapanya, S., & Singh, S. (2005). Synchrony in the onset of mental-state reasoning: Evidence from five cultures. *Psychological Science*, *16*(5), 378–384.

Campos, B., Graesch, A. P., Repetti, R., Bradbury, T., & Ochs, E. (2009). Opportunity for interaction? A naturalistic observation study of dual-earner families after work and school. *Journal of Family Psychology*, *23*(6), 798.

Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., & Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development*, i–174.

Carpenter, M., & Tomasello, M. (1995). Joint attention and imitative learning in children, chimpanzees, and enculturated chimpanzees. *Social Development*, *4*(3), 217–237.

Carr, A., Slade, L., Yuill, N., Sullivan, S., & Ruffman, T. (2018). Minding the children: A longitudinal study of mental state talk, theory of mind, and behavioural adjustment from

the age of 3 to 10. *Social Development*, *27*(4), 826–840. https://doi.org/10.1111/sode.12315

Carston, R. (2004). Relevance theory and the saying/implicating distinction. *The Handbook of Pragmatics*, 633–656.

Carter, C. S. (2014). Oxytocin pathways and the evolution of human behavior. *Annual Review of Psychology*, *65*, 17–39.

Casasanto, D. (2015). Linguistic relativity. *The Routledge Handbook of Semantics*, 158.

Castelli, F., Frith, C., Happé, F., & Frith, U. (2002). Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain*, *125*(8), 1839–1849.

Castelli, F., Happé, F., Frith, U., & Frith, C. (2000). Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns. *Neuroimage*, *12*(3), 314–325.

Chahuneau, V., Schlinger, E., Smith, N. A., & Dyer, C. (n.d.). *Translating into Morphologically Rich Languages with Synthetic Phrases*. 11.

Chang, M.-J. (2009). *A cross-cultural study of Taiwanese and British university students' oral narratives*. https://figshare.le.ac.uk/articles/thesis/A_cross-cultural_study_of_Taiwanese_and_British_university_students_oral_narratives/10093133

Chen, X. (1985). The one-child population policy, modernization, and the extended Chinese family. *Journal of Marriage and the Family*, 193–202.

Cheney, D., Seyfarth, R., & Smuts, B. (1986). Social relationships and social cognition in nonhuman primates. *Science*, *234*(4782), 1361–1366.

Cheng, J. T., Tracy, J. L., Foulsham, T., Kingstone, A., & Henrich, J. (2013). Two ways to the top: Evidence that dominance and prestige are distinct yet viable avenues to social rank and influence. *Journal of Personality and Social Psychology*, *104*(1), 103.

Cheung, H., Chen, H.-C., & Yeung, W. (2009). Relations between mental verb and false belief understanding in Cantonese-speaking children. *Journal of Experimental Child Psychology*, *104*(2), 141–155.

Cheung, H., Hsuan-Chih, C., Creed, N., Ng, L., Ping Wang, S., & Mo, L. (2004). Relative Roles of General and Complementation Language in Theory-of-Mind Development: Evidence From Cantonese and English. *Child Development*, *75*(4), 1155–1170. https://doi.org/10.1111/j.1467-8624.2004.00731.x

Chevallier, C., Kohls, G., Troiani, V., Brodkin, E. S., & Schultz, R. T. (2012). The social motivation theory of autism. *Trends in Cognitive Sciences*, *16*(4), 231–239.

China Statistical Yearbook. (2022). National Bureau of Statistics of China. https://www.stats.gov.cn/sj/ndsj/2022/indexeh.htm

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT press.

Christiansen, M. H., & Kirby, S. (2003). Language evolution: Consensus and controversies. *Trends in Cognitive Sciences*, *7*(7), 300–307.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204.

Clark, H. (1996). *Using language. New York, NY, US*. Cambridge University Press. http://dx. doi. org/10.2277/0521561582.

Cochet, H., Jover, M., Rizzo, C., & Vauclair, J. (2017). Relationships between declarative pointing and theory of mind abilities in 3-to 4-year-olds. *European Journal of Developmental Psychology*, *14*(3), 324–336.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Cohn, N., Taylor-Weiner, A., & Grossman, S. (2012). Framing Attention in Japanese and American Comics: Cross-Cultural Differences in Attentional Structure. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00349

Collins, J. A., & Olson, I. R. (2014). Knowledge is power: How conceptual knowledge transforms visual cognition. *Psychonomic Bulletin & Review*, *21*, 843–860.

Conley, R. (2015). *Confronting the death penalty: How language influences jurors in capital cases*. Oxford University Press.

Csibra, G., & Gergely, G. (2006). Social learning and social cognition: The case for pedagogy. *Processes of Change in Brain and Cognitive Development. Attention and Performance XXI*, *21*, 249–274.

Dale, R., & Lupyan, G. (2012). Understanding the origins of morphological diversity: The linguistic niche hypothesis. *Advances in Complex Systems*, *15*(03n04), 1150017.

Daly, M., Wilson, M., & Weghorst, S. J. (1982). Male sexual jealousy. *Ethology and Sociobiology*, *3*(1), 11–27.

Dasen, P. R., & Mishra, R. C. (2010). *Development of geocentric spatial language and cognition: An eco-cultural perspective* (Vol. 12). Cambridge University Press.

Davies, J. R., & Garcia-Pelegrin, E. (2023). Bottlenose dolphins are sensitive to human attentional features, including eye functionality. *Scientific Reports*, *13*(1), 12565. https://doi.org/10.1038/s41598-023-39031-7

De Rosnay, M., Fink, E., Begeer, S., Slaughter, V., & Peterson, C. (2014). Talking theory of mind talk: Young school-aged children's everyday conversation and understanding of mind and emotion. *Journal of Child Language*, *41*(5), 1179–1193.

de Villiers, J. G. (2005). Can Language Acquisition Give Children a Point of View? In *Why Language Matters for Theory of Mind*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195159912.003.0010

de Villiers, J. G., & Pyers, J. E. (2002). Complements to cognition: A longitudinal study of the relationship between complex syntax and false-belief-understanding. *Cognitive Development*, *17*(1), 1037–1060. https://doi.org/10.1016/S0885-2014(02)00073-4

Dennett, D. C. (1978). Beliefs about beliefs. *Behavioral and Brain Sciences*, *1*(4), 568–570.

D'Entremont, B., Hains, S. M., & Muir, D. W. (1997). A demonstration of gaze following in 3-to 6-month-olds. *Infant Behavior and Development*, *20*(4), 569–572.

Devine, R. T., & Hughes, C. (2019). Let's Talk: Parents' Mental Talk (Not Mind-Mindedness or Mindreading Capacity) Predicts Children's False Belief Understanding. *Child Development*, *90*(4), 1236–1253.

Dixson, H. G., Komugabe-Dixson, A. F., Dixson, B. J., & Low, J. (2017). Scaling Theory of Mind in a Small-Scale Society: A Case Study From Vanuatu. *Child Development*.

Doody, J. S., Burghardt, G. M., & Dinets, V. (2013). Breaking the Social–Non-social Dichotomy: A Role for Reptiles in Vertebrate Social Behavior Research? *Ethology*, *119*(2), 95–103. https://doi.org/10.1111/eth.12047

Dubey, I., Ropar, D., & de C Hamilton, A. F. (2015). Measuring the value of social engagement in adults with and without autism. *Molecular Autism*, *6*, 1–9.

Dunbar, R. I. (2004). Gossip in evolutionary perspective. *Review of General Psychology*, *8*(2), 100–110.

Dunning, D., & Balcetis, E. (2013). Wishful seeing: How preferences shape visual perception. *Current Directions in Psychological Science*, *22*(1), 33–37.

Duranti, A. (2008). Further reflections on reading other minds. *Anthropological Quarterly*, *81*(2), 483–494.

Durrleman, S., Burnel, M., De Villiers, J. G., Thommen, E., Yan, R., & Delage, H. (2019). The impact of grammar on mentalizing: A training study including children with autism spectrum disorder and developmental language disorder. *Frontiers in Psychology*, *10*, 2478.

Dutemple, E., & Sheldon, S. (2022). The effect of retrieval goals on the content recalled from complex narratives. *Memory & Cognition*, *50*(2), 397–406.

Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., Kessler, J., Woike, J. K., Wolf, O. T., & Convit, A. (2006). Introducing MASC: a movie for the assessment of social cognition. *Journal of Autism and Developmental Disorders*, *36*, 623–636.

Ekman, P. (1992). Are there basic emotions?. *Psychological Review*, *99*(3), 550.

Ekman, P., Friesen, W. V., O'sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W. A., Pitcairn, T., Ricci-Bitti, P. E., & others. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, *53*(4), 712.

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*, *128*(2), 203.

Emery, N. J., & Clayton, N. S. (2001). Effects of experience and social context on prospective caching strategies by scrub jays. *Nature*, *414*(6862), 443.

Ezeizabarrena, M.-J., & Garcia Fernandez, I. (2018). Length of utterance, in morphemes or in words?: MLU3-w, a reliable measure of language development in early basque. *Frontiers in Psychology*, *8*, 2265.

Fargues, P. (2011). International migration and the demographic transition: A two-way interaction. *International Migration Review*, *45*(3), 588–614.

Farroni, T., Csibra, G., Simion, F., & Johnson, M. H. (2002). Eye contact detection in humans from birth. *Proceedings of the National Academy of Sciences*, *99*(14), 9602–9605.

Fausey, C. M., & Boroditsky, L. (2008). English and Spanish speakers remember causal agents differently. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *30*.

Fausey, C. M., & Boroditsky, L. (2010). Subtle linguistic cues influence perceived blame and financial liability. *Psychonomic Bulletin & Review*, *17*(5), 644–650.

Fausey, C. M., & Boroditsky, L. (2011). Who dunnit? Cross-linguistic differences in eye-witness memory. *Psychonomic Bulletin & Review*, *18*(1), 150–157. https://doi.org/10.3758/s13423-010-0021-5

Fausey, C. M., Long, B. L., & Boroditsky, L. (2009). The role of language in eyewitness memory: Remembering who did it in English and Japanese. *Manuscript in Preparation*. http://csjarchive.cogsci.rpi.edu/Proceedings/2009/papers/559/paper559.pdf

Fehr, B., & Russell, J. A. (1984). Concept of emotion viewed from a prototype perspective. *Journal of Experimental Psychology: General*, *113*(3), 464–486. https://doi.org/10.1037/0096-3445.113.3.464

Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature*, *13*(1), 1–25.

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for “top-down” effects. *Behavioral and Brain Sciences*, *39*, e229. https://doi.org/10.1017/S0140525X15000965

Fitch, W. T., Hauser, M. D., & Chomsky, N. (2005). The evolution of the language faculty: Clarifications and implications. *Cognition*, *97*(2), 179–210.

Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children’s knowledge about visual perception: Further evidence for the Level 1–Level 2 distinction. *Developmental Psychology*, *17*(1), 99.

Floyd, S., Rossi, G., Baranova, J., Blythe, J., Dingemanse, M., Kendrick, K. H., Zinken, J., & Enfield, N. J. (2018). Universals and cultural diversity in the expression of gratitude. *Royal Society Open Science*, *5*(5), 180391.

Fodor, J. (1975). *The language of thought*. Harvard University Press.

Fodor, J. (1981). Propositional attitudes. In *The Language and Thought Series* (pp. 45–63). Harvard University Press.

Fodor, J. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. MIT Press.

Fodor, J. (1992). A theory of the child's theory of mind. *Cognition*.

Frank, M. C., Tenenbaum, J. B., & Fernald, A. (2013). Social and Discourse Contributions to the Determination of Reference in Cross-Situational Word Learning. *Language Learning and Development*, *9*(1), 1–24. https://doi.org/10.1080/15475441.2012.707101

Frederickx, S., & Hofmans, J. (2014). The role of personality in the initiation of communication situations. *Journal of Individual Differences*.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, *2*(12), 493–501.

Gendron, M., Roberson, D., van der Vyver, J. M., & Barrett, L. F. (2014). Perceptions of emotion from facial expressions are not culturally universal: Evidence from a remote culture. *Emotion*, *14*(2), 251.

Geography of Philosophy Project. (2017, December 7). About the Project. *Geography of Philosophy*. https://www.geographyofphilosophy.com/about

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*(2), 165–193.

Giannakopoulou, L., Kavouras, M., Kokla, M., & Mark, D. (2013). From compasses and maps to mountains and territories: Experimental results on geographic cognitive categorization.

In *Cognitive and Linguistic Aspects of Geographic Space* (pp. 63–81). Springer. http://link.springer.com/chapter/10.1007/978-3-642-34359-9_4

Gilbert, S. F., Bosch, T. C., & Ledón-Rettig, C. (2015). Eco-Evo-Devo: Developmental symbiosis and developmental plasticity as evolutionary agents. *Nature Reviews Genetics*, *16*(10), 611.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*(1), 3–55.

Goddard, C. (2010). *Universals and Variation in the Lexicon of Mental State Concepts*. Oxford University Press. http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780195311129.001.0001/acprof-9780195311129-chapter-5

Goddard, C., & Wierzbicka, A. (1994). *Semantic and lexical universals: Theory and empirical findings*.

Goddard, C., & Wierzbicka, A. (2002). *Meaning and universal grammar: Theory and empirical findings* (Vol. 1). John Benjamins Publishing.

Golan, O., Baron-Cohen, S., Hill, J. J., & Rutherford, M. (2007). The 'Reading the Mind in the Voice'test-revised: A study of complex emotion recognition in adults with and without autism spectrum conditions. *Journal of Autism and Developmental Disorders*, *37*(6), 1096–1106.

Goldstein, T. R., & Winner, E. (2011). Engagement in role play, pretense, and acting classes predict advanced theory of mind skill in middle childhood. *Imagination, Cognition and Personality*, *30*(3), 249–258.

Goldstone, R. L., de Leeuw, J. R., & Landy, D. H. (2015). Fitting perception in and to cognition. *Cognition*, *135*, 24–29.

Gonzales, A. L., Hancock, J. T., & Pennebaker, J. W. (2010). Language style matching as a predictor of social dynamics in small groups. *Communication Research*, *37*(1), 3–19.

Gopnik, A., & Astington, J. W. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Development*, 26–37.

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind & Language*, *7*(1–2), 145–171.

Gorman, B. K., Fiestas, C. E., Peña, E. D., & Clark, M. R. (2011). Creative and Stylistic Devices Employed by Children During a Storybook Narrative Task: A Cross-Cultural Study. *Language, Speech, and Hearing Services in Schools*, *42*(2), 167–181. https://doi.org/10.1044/0161-1461(2010/10-0052)

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619–619.

Greenberg, D. M., Warrier, V., Abu-Akel, A., Allison, C., Gajos, K. Z., Reinecke, K., Rentfrow, P. J., Radecki, M. A., & Baron-Cohen, S. (2023). Sex and age differences in "theory of mind" across 57 countries using the English version of the "Reading the Mind in the Eyes" Test. *Proceedings of the National Academy of Sciences*, *120*(1), e2022385119. https://doi.org/10.1073/pnas.2022385119

Greenberg, J. H. (Ed.). (1963). *Universals of language.* M.I.T. Press.

Grice, H. P. (1975). Logic and conversation. *1975*, 41–58.

Gumperz, J. J., & Levinson, S. C. (1991). Rethinking linguistic relativity. *Current Anthropology*, *32*(5), 613–623.

Gumperz, J. J., & Levinson, S. C. (1996). *Rethinking linguistic relativity*. Cambridge University Press.

Hall, E. T. (1973). *The silent language*. Anchor.

Hall, W. S., Nagy, W. E., & Nottenburg, G. (1981). Situational variation in the use of internal state words. *Center for the Study of Reading Technical Report; No. 212*.

Hansen, N., Porter, J. D., & Francis, K. (n.d.). *A Corpus Study of "Know": On The Verification of Philosophers' Frequency Claims about Language*. 34.

Happé, F. G. (1993). Communicative competence and theory of mind in autism: A test of relevance theory. *Cognition*, *48*(2), 101–119.

Hare, B., Call, J., Agnetta, B., & Tomasello, M. (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behaviour*, *59*(4), 771–785.

Hargreaves, D. (2005). Agency and intentional action in Kathmandu Newar. *Himalayan Linguistics*, *5*, 1–48.

Harrigan, K., Hacquard, V., & Lidz, J. (2018). Three-Year-Olds' Understanding of Desire Reports Is Robust to Conflict. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00119

Haselton, M. G., & Buss, D. M. (2000). Error management theory: A new perspective on biases in cross-sex mind reading. *Journal of Personality and Social Psychology*, *78*(1), 81.

Haspelmath, M. (2010). Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, *86*(3), 663–687. https://doi.org/10.1353/lan.2010.0021

Haun, D. B., Rapold, C. J., Janzen, G., & Levinson, S. C. (2011). Plasticity of human spatial cognition: Spatial language and cognition covary across cultures. *Cognition*, *119*(1), 70–80.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, *298*(5598), 1569–1579.

Hawkins, R. X., & Goodman, N. D. (2016). Conversational expectations account for apparent limits on theory of mind use. *Proceedings of the Thirty-Eighth Annual Conference of the Cognitive Science Society*, 1889–1894.

Heine, B. (1997). *Cognitive foundations of grammar*. Oxford University Press.

Henrich, J., & Gil-White, F. J. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behavior*, *22*(3), 165–196.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hevia, M. D. de, Izard, V., Coubart, A., Spelke, E. S., & Streri, A. (2014). Representations of space, time, and number in neonates. *Proceedings of the National Academy of Sciences*, *111*(13), 4809–4813. https://doi.org/10.1073/pnas.1323628111

Heyes, C. (2018). *Cognitive Gadgets: The Cultural Evolution of Thinking*. Harvard University Press. https://books.google.com/books?id=lbpTDwAAQBAJ

Hoenigman, D. (2015). *"The talk goes many ways": Registers of language and modes of performance in Kanjimei, East Sepik Province, Papua New Guinea*. https://doi.org/10.25911/5c4834f04778a

Hoffman, C., Lau, I., & Johnson, D. R. (1986). The linguistic relativity of person cognition: An English–Chinese comparison. *Journal of Personality and Social Psychology*, *51*(6), 1097.

Hoijer, H. E. (1954). *Language in culture; conference on the interrelations of language and other aspects of culture.*

Holekamp, K. E. (2007). Questioning the social intelligence hypothesis. *Trends in Cognitive Sciences*, *11*(2), 65–69.

Huang, M., & Jaszczolt, K. M. (2018). *Expressing the Self: Cultural Diversity and Cognitive Universals*. Oxford University Press.

Huber, L., & Lonardo, L. (2023). Canine perspective-taking. *Animal Cognition*, *26*(1), 275–298. https://doi.org/10.1007/s10071-022-01736-z

Huettig, F., Chen, J., Bowerman, M., & Majid, A. (2010). Do language-specific categories shape conceptual processing? Mandarin classifier distinctions influence eye gaze behavior, but only during linguistic processing. *Journal of Cognition and Culture*, *10*(1), 39–58. https://doi.org/10.1163/156853710X497167

Hughes, C., & Devine, R. T. (2015). Individual Differences in Theory of Mind From Preschool to Adolescence: Achievements and Directions. *Child Development Perspectives*, *9*(3), 149–153. https://doi.org/10.1111/cdep.12124

Hughes, C., Devine, R. T., Ensor, R., Koyasu, M., Mizokawa, A., & Lecce, S. (2014). Lost in translation? Comparing British, Japanese, and Italian children□s theory-of-mind performance. *Child Development Research*, *2014*.

Hughes, C., Devine, R. T., & Wang, Z. (2018). Does parental mind-mindedness account for cross-cultural differences in preschoolers□ theory of mind? *Child Development*, *89*(4), 1296–1310.

Hughes, C., & Dunn, J. (1998). Understanding mind and emotion: Longitudinal associations with mental-state talk between young friends. *Developmental Psychology*, *34*(5), 1026.

Imai, M., & Gentner, D. (1997). A cross-linguistic study of early word meaning: Universal ontology and linguistic influence. *Cognition*, *62*(2), 169–200.

Jackson, J. C., Watts, J., Henry, T. R., List, J.-M., Forkel, R., Mucha, P. J., Greenhill, S. J., Gray, R. D., & Lindquist, K. A. (2019). Emotion semantics show both cultural variation and universal structure. *Science*, *366*(6472), 1517–1522.

Johnson, C. (1999). Metaphor vs. Conflation in the Acquisition of Polysemy: The Case of SEE.''. *Cultural, Psychological and Typological Issues in Cognitive Linguistics: Selected Papers of the Bi-Annual ICLA Meeting in Albuquerque, July 1995*, *152*, 155. https://books.google.nl/books?hl=en&lr=&id=lWBCAAAAQBAJ&oi=fnd&pg=PA155&dq=c

hristopher+johnson+metaphor+conflation&ots=94lQZII1Vo&sig=H2cycsJKBU9evMX4p8PHZnxegNA

Jolliffe, T., & Baron-Cohen, S. (1999). The strange stories test: A replication with high-functioning adults with autism or Asperger syndrome. *Journal of Autism and Developmental Disorders*, *29*, 395–406.

Kano, F., Krupenye, C., Hirata, S., Call, J., & Tomasello, M. (2017). Submentalizing cannot explain belief-based action anticipation in apes. *Trends in Cognitive Sciences*, *21*(9), 633–634.

Karmiloff-Smith, A., Klima, E., Bellugi, U., Grant, J., & Baron-Cohen, S. (1995). Is There a Social Module? Language, Face Processing, and Theory of Mind in Individuals with Williams Syndrome. *Journal of Cognitive Neuroscience*, *7*(2), 196–208. https://doi.org/10.1162/jocn.1995.7.2.196

Kay, P., & Regier, T. (2003). Resolving the question of color naming universals. *Proceedings of the National Academy of Sciences*, *100*(15), 9085–9089.

Kay, P., & Regier, T. (2006). Language, thought and color: Recent developments. *Trends in Cognitive Sciences*, *10*(2), 51–54.

Kim, H., Kaduthodil, J., Strong, R. W., Germine, L., Cohan, S., & Wilmer, J. B. (2022). *Multiracial Reading the Mind in the Eyes Test (MRMET): An inclusive version of an influential measure*.

Kim, Y.-S. G., Dore, R., Cho, M., Golinkoff, R., & Amendum, S. J. (2021). Theory of mind, mental state talk, and discourse comprehension: Theory of mind process is more important for narrative comprehension than for informational text comprehension. *Journal of Experimental Child Psychology*, *209*, 105181. https://doi.org/10.1016/j.jecp.2021.105181

Kockelman, P. (2006). Representations of the world: Memories, perceptions, beliefs, intentions, and plans. *Semiotica*, *2006*(162), 73–125.

Koster-Hale, J., & Saxe, R. (2013). Theory of mind: A neural prediction problem. *Neuron*, *79*(5), 836–848.

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, *330*(6012), 1830–1834.

Kristen, S., Chiarella, S., Sodian, B., Aureli, T., Genco, M., & Poulin-Dubois, D. (2014). *Crosslinguistic Developmental Consistency in the Composition of Toddlers' Internal State Vocabulary: Evidence from Four Languages* [Research article]. Child Development Research. https://doi.org/10.1155/2014/575142

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, *354*(6308), 110–114.

Kulke, L., Johannsen, J., & Rakoczy, H. (2019). Why can some implicit Theory of Mind tasks be replicated and others cannot? A test of mentalizing versus submentalizing accounts. *PloS One*, *14*(3).

Kuntoro, I. A., Saraswati, L., Peterson, C., & Slaughter, V. (2013). Micro-cultural influences on theory of mind development: A comparative study of middle-class and pemulung children in Jakarta, Indonesia. *International Journal of Behavioral Development*, *37*(3), 266–273.

Kwisthout, J., Vogt, P., Haselager, P., & Dijkstra, T. (2008). Joint attention and language evolution. *Connection Science*, *20*(2–3), 155–171.

Lagattuta, K. H., Sayfan, L., & Blattman, A. J. (2010). Forgetting common ground: Six-to seven-year-olds have an overinterpretive theory of mind. *Developmental Psychology*, *46*(6), 1417.

Lakoff, G. (2008). *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago press.

Laroui, A., Brown, L. C., Barbour, N., Miller, S. G., & Swearingen, W. D. (2024). Morocco. In *Encyclopedia Brittanica*. https://www.britannica.com/place/Morocco

Lecce, S., Ronchi, L., & Devine, R. T. (2021). Mind what teacher says: Teachers' propensity for mental-state language and children's theory of mind in middle childhood. *Social Development*.

Lehrner, J., Glück, J., & Laska, M. (1999). Odor identification, consistency of label use, olfactory threshold and their relationships to odor memory over the human lifespan. *Chemical Senses*, *24*(3), 337–346. https://doi.org/10.1093/chemse/24.3.337

Leslie, A. M. (1987). Pretense and representation: The origins of" theory of mind.". *Psychological Review*, *94*(4), 412.

Leslie, A. M. (1994). ToMM, ToBy, and Agency: Core architecture and domain specificity. *Mapping the Mind: Domain Specificity in Cognition and Culture*, 119–148.

Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in theory of mind. *Trends in Cognitive Sciences*, *8*(12), 528–533.

Leslie, A. M., & Happé, F. (1989). Autism and ostensive communication: The relevance of metarepresentation. *Development and Psychopathology*, *1*(3), 205–212.

Levinson, P., Brown, P., Levinson, S. C., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge university press.

Levis, N. A., & Pfennig, D. W. (2016). Evaluating 'plasticity-first'evolution in nature: Key criteria and empirical approaches. *Trends in Ecology & Evolution*, *31*(7), 563–574.

Li, P. (1993). Cryptotypes, meaning-form mappings, and overgeneralizations. *Proceedings of the Twenty-Fourth Annual Child Language Research Form. Stanford: Center for the Study of Language and Information*.

Liebal, K., Behne, T., Carpenter, M., & Tomasello, M. (2009). Infants use shared experience to interpret pointing gestures. *Developmental Science*, *12*(2), 264–271.

Lillard, A. (1998). Ethnopsychologies: Cultural variations in theories of mind. *Psychological Bulletin*, *123*(1), 3.

Lindström, A., & Sorjonen, M.-L. (2012). Affiliation in conversation. *The Handbook of Conversation Analysis*, 250–369.

Lindström, J., & Karlsson, S. (2016). Tensions in the epistemic domain and claims of no-knowledge: A study of Swedish medical interaction. *Journal of Pragmatics*, *106*, 129–147. https://doi.org/10.1016/j.pragma.2016.07.003

Liu, D., Wellman, H. M., Tardif, T., & Sabbagh, M. A. (2008). Theory of mind development in Chinese children: A meta-analysis of false-belief understanding across cultures and languages. *Developmental Psychology*, *44*(2), 523.

Lo, R. F., & Mar, R. A. (2022). Having siblings is associated with better mentalizing abilities in adults. *Cognitive Development*, *63*, 101193. https://doi.org/10.1016/j.cogdev.2022.101193

Lupyan, G. (2012). Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in Psychology*, *3*, 54.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PloS One*, *5*(1), e8559.

Lyons, D. E., & Santos, L. R. (2006). Ecology, domain specificity, and the origins of theory of mind: Is competition the catalyst? *Philosophy Compass*, *1*(5), 481–492.

Lyons, D. E., Young, A. G., & Keil, F. C. (2007). The hidden structure of overimitation. *Proceedings of the National Academy of Sciences*, *104*(50), 19751–19756.

Majid, A., Bowerman, M., Staden, M. van, & Boster, J. S. (2007). The semantic categories of cutting and breaking events: A crosslinguistic perspective. *Cognitive Linguistics*, *18*(2), 133–152. https://doi.org/10.1515/COG.2007.005

Majid, A., & Burenhult, N. (2014). Odors are expressible in language, as long as you speak the right language. *Cognition*, *130*(2), 266–270. https://doi.org/10.1016/j.cognition.2013.11.004

Majid, A., Roberts, S. G., Cilissen, L., Emmorey, K., Nicodemus, B., O'grady, L., Woll, B., LeLan, B., De Sousa, H., Cansler, B. L., & others. (2018). Differential coding of perception in the world's languages. *Proceedings of the National Academy of Sciences*, *115*(45), 11369–11376.

Matsumoto, D. (1989). Cultural influences on the perception of emotion. *Journal of Cross-Cultural Psychology*, *20*(1), 92–105.

McAlister, A., & Peterson, C. (2007). A longitudinal study of child siblings and theory of mind development. *Cognitive Development*, *22*(2), 258–270. https://doi.org/10.1016/j.cogdev.2006.10.009

McCabe, K. A., Smith, V. L., & LePore, M. (2000). Intentionality detection and "mindreading": Why does game form matter? *Proceedings of the National Academy of Sciences*, *97*(8), 4404–4409.

McComiskey, B. (2002). *Gorgias and the new sophistic rhetoric*. SIU Press.

McCroskey, J. C., & Richmond, V. P. (1995). Correlates of compulsive communication: Quantitative and qualitative characteristics. *Communication Quarterly*, *43*(1), 39–52.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282.

Mehl, M. R., Vazire, S., Ramírez-Esparza, N., Slatcher, R. B., & Pennebaker, J. W. (2007). Are women really more talkative than men? *Science*, *317*(5834), 82–82.

Meins, E., Fernyhough, C., & Harris-Waller, J. (2014). Is mind-mindedness trait-like or a quality of close relationships? Evidence from descriptions of significant others, famous people, and works of art. *Cognition*, *130*(3), 417–427.

Méndez, L. I., Bitetti, D., & Perry, J. (2023). A cross-cultural perspective of narrative retells in kindergarten children. *Bilingual Research Journal*, *46*(3–4), 275–289. https://doi.org/10.1080/15235882.2023.2258838

Migge, B., & Léglise, I. (2007). 10. Language and colonialism. In M. Hellinger & A. Pauwels (Eds.), *Handbook of Language and Communication: Diversity and Change* (pp. 299–332). De Gruyter Mouton. https://doi.org/doi:10.1515/9783110198539.2.299

Miller, J. G. (1986). Early cross-cultural commonalities in social explanation. *Developmental Psychology*, *22*(4), 514.

Miller, S. A. (2009). Children's understanding of second-order mental states. *Psychological Bulletin*, *135*(5), 749.

Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Development*, *78*(2), 622–646.

Milligan, L. (2016). Insider-outsider-inbetweener? Researcher positioning, participative methods and cross-cultural educational research. *Compare: A Journal of Comparative and International Education*, *46*(2), 235–250.

Misyak, J., Noguchi, T., & Chater, N. (2016). Instantaneous conventions: The emergence of flexible communicative signals. *Psychological Science*, *27*(12), 1550–1561.

Moll, H., & Tomasello, M. (2007). Cooperation and human cognition: The Vygotskian intelligence hypothesis. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *362*(1480), 639–648.

Moya, C., & Henrich, J. (2016). Culture–gene coevolutionary psychology: Cultural learning, language, and ethnic psychology. *Current Opinion in Psychology*, *8*, 112–118.

Munnich, E., Landau, B., & Dosher, B. A. (2001). Spatial language and spatial representation: A cross-linguistic comparison. *Cognition*, *81*(3), 171–208. https://doi.org/10.1016/S0010-0277(01)00127-5

Murtha, T. C., Kanfer, R., & Ackerman, P. L. (1996). Toward an interactionist taxonomy of personality and situations: An integrative situational—Dispositional representation of personality traits. *Journal of Personality and Social Psychology*, *71*(1), 193.

Nelson, M. (2023). Propositional Attitude Reports. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2023). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2023/entries/prop-attitude-reports/

Nettle, D. (2012). Social scale and structural complexity in human languages. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1597), 1829–1836.

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, *45*(3), 211–236.

Nichols, S., & Stich, S. P. (2003). *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds.* Clarendon Press/Oxford University Press.

Nilsson, K. K., & de López, K. J. (2016). Theory of Mind in Children With Specific Language Impairment: A Systematic Review and Meta-Analysis. *Child Development*, *87*(1), 143–153. https://doi.org/10.1111/cdev.12462

Norenzayan, A., & Heine, S. J. (2005). Psychological universals: What are they and how can we know? *Psychological Bulletin*, *131*(5), 763.

Noyes, A. F. (1944). Early causes and development of the doctrine of mens rea. *Ky. LJ*, *33*, 306.

Ochs, E., Shohet, M., Campos, B., & Beck, M. (2011). 3. Coming Together at Dinner: A Study of Working Families. In K. Christensen & B. Schneider (Eds.), *Realigning 20th-Century*

Jobs for a 21st-Century Workforce (pp. 57–70). Cornell University Press. https://doi.org/doi:10.7591/9780801458446-006

Okanda, M., Asada, K., Moriguchi, Y., & Itakura, S. (2015). Understanding violations of Gricean maxims in preschoolers and adults. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00901

Oleszkiewicz, A., Walliczek-Dworschak, U., Klötze, P., Gerber, F., Croy, I., & Hummel, T. (2016). Developmental Changes in Adolescents' Olfactory Performance and Significance of Olfaction. *PLOS ONE*, *11*(6), e0157560. https://doi.org/10.1371/journal.pone.0157560

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*(5719), 255–258.

Paal, T., & Bereczkei, T. (2007). Adult theory of mind, cooperation, Machiavellianism: The effect of mindreading on social relations. *Personality and Individual Differences*, *43*(3), 541–551.

Papafragou, A., Cassidy, K., & Gleitman, L. (2007). When we think about thinking: The acquisition of belief verbs. *Cognition*, *105*(1), 125–165. https://doi.org/10.1016/j.cognition.2006.09.008

Papafragou, A., & Li, P. (2001). Evidential morphology and theory of mind. *Proceedings from the 26th Annual Boston University Conference on Language Development. Cascadilla Press, Somerville, MA*, 510–520.

Parrigon, S., Woo, S. E., Tay, L., & Wang, T. (2017). CAPTION-ing the situation: A lexically-derived taxonomy of psychological situation characteristics. *Journal of Personality and Social Psychology*, *112*(4), 642.

Passban, P. (2017). *Machine translation of morphologically rich languages using deep neural networks* [Doctoral, Dublin City University]. http://doras.dcu.ie/22200/

Penn, J. M. (2014). *Linguistic relativity versus innate ideas: The origins of the Sapir-Whorf hypothesis in German thought* (Vol. 120). Walter de Gruyter.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, *71*(2001), 2001.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, *77*(6), 1296.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, *54*(1), 547–577.

Perez-Zapata, D., Slaughter, V., & Henry, J. D. (2016). Cultural effects on mindreading. *Cognition*, *146*, 410–414.

Perner, J. (1988). Developing semantics for theories of mind: From propositional attitudes to mental representation. *Developing Theories of Mind*, 141–172.

Perner, J., Ruffman, T., & Leekam, S. R. (1994). Theory of mind is contagious: You catch it from your sibs. *Child Development*, *65*(4), 1228–1238.

Perner, J., Sprung, M., Zauner, P., & Haider, H. (2003). Want That is Understood Well before Say That, Think That, and False Belief: A Test of de Villiers's Linguistic Determinism on German–Speaking Children. *Child Development*, *74*(1), 179–188. https://doi.org/10.1111/1467-8624.t01-1-00529

Persson, T., Sauciuc, G.-A., Fantasia, V., & Bard, K. (2023). Editorial: Exploring shared intentionality: Underlying mechanisms, evolutionary roots, developmental trajectories, and cultural influences. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1332592

Phillips, W., & Boroditsky, L. (2003). Can quirks of grammar affect the way you think? Grammatical gender and object concepts. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *25*(25).

Pinker, S. (2003). *The language instinct: How the mind creates language*. Penguin uK.

Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences*, *13*(4), 707–727.

Pinto, G., Primi, C., Tarchi, C., & Bigozzi, L. (2017). *Mental State Talk Structure in Children's Narratives: A Cluster Analysis* [Research article]. Child Development Research. https://doi.org/10.1155/2017/1725487

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526.

Prewitt-Freilino, J. L., Caswell, T. A., & Laakso, E. K. (2012). The Gendering of Language: A Comparison of Gender Equality in Countries with Gendered, Natural Gender, and Genderless Languages. *Sex Roles*, *66*(3–4), 268–281. https://doi.org/10.1007/s11199-011-0083-5

Proost, K. (2007). *Conceptual Structure in Lexical Items: The lexicalisation of communication concepts in English, German and Dutch*. John Benjamins Publishing.

PRRI. (2021). The 2020 census of American religion. *Public Religion Research Institute*, *7*.

Quaranta, A., d'Ingeo, S., Amoruso, R., & Siniscalchi, M. (2020). Emotion Recognition in Cats. *Animals*, *10*(7). https://doi.org/10.3390/ani10071107

Quesque, F., & Rossetti, Y. (2020). What do theory-of-mind tasks actually measure? Theory and practice. *Perspectives on Psychological Science*, *15*(2), 384–396.

Range, F., & Virányi, Z. (2011). Development of gaze following abilities in wolves (Canis lupus). *PLoS One*, *6*(2), e16888.

Rauthmann, J. F., Gallardo-Pujol, D., Guillaume, E. M., Todd, E., Nave, C. S., Sherman, R. A., Ziegler, M., Jones, A. B., & Funder, D. C. (2014). The Situational Eight DIAMONDS: A taxonomy of major dimensions of situation characteristics. *Journal of Personality and Social Psychology*, *107*(4), 677.

Regier, T., & Kay, P. (2009). Language, thought, and color: Whorf was half right. *Trends in Cognitive Sciences*, *13*(10), 439–446.

Richerson, P., Baldini, R., Bell, A. V., Demps, K., Frost, K., Hillis, V., Mathew, S., Newton, E. K., Naar, N., Newson, L., & others. (2016). Cultural group selection plays an essential role in explaining human cooperation: A sketch of the evidence. *Behavioral and Brain Sciences*, *39*.

Robbins, J. (2004). *Becoming Sinners: Christianity and Moral Torment in a Papua New Guinea Society* (1st ed.). University of California Press. http://www.jstor.org/stable/10.1525/j.ctt1pp8f0

Robbins, J., & Rumsey, A. (2008). Introduction: Cultural and linguistic anthropology and the opacity of other minds. *Anthropological Quarterly*, *81*(2), 407–420.

Roby, E., & Scott, R. M. (2022). Exploring the impact of parental education, ethnicity and context on parent and child mental-state language. *Cognitive Development*, *62*, 101169. https://doi.org/10.1016/j.cogdev.2022.101169

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439.

Ruffman, T., Slade, L., & Crowe, E. (2002). The Relation between Children's and Mothers' Mental State Language and Theory-of-Mind Understanding. *Child Development*, *73*(3), 734–751. https://doi.org/10.1111/1467-8624.00435

Ruggles, S. (1994). The transformation of American family structure. *The American Historical Review*, *99*(1), 103–128.

Rutherford, M. D., Baron-Cohen, S., & Wheelwright, S. (2002). Reading the mind in the voice: A study with normal adults and adults with Asperger syndrome and high functioning autism. *Journal of Autism and Developmental Disorders*, *32*(3), 189–194.

Saalbach, H., & Imai, M. (2011). The relation between linguistic categories and cognition: The case of numeral classifiers. *Language and Cognitive Processes*, 1–48. https://doi.org/10.1080/01690965.2010.546585

Said, E. W. (2016). *Orientalism: Western conceptions of the Orient*. Penguin UK.

Sally, D., & Hill, E. (2006). The development of interpersonal strategy: Autism, theory-of-mind, cooperation and fairness. *Journal of Economic Psychology*, *27*(1), 73–97.

Salmond, A. (n.d.). Theoretical Landscapes: On Cross-cultural conceptions of knowledge. *Ed. David Parkin, Semantic Anthropology, ASA Monograph Series No.22, (London, Academic Press),*. Retrieved June 13, 2019, from https://www.academia.edu/39528388/Theoretical_Landscapes_On_Cross-cultural_conceptions_of_knowledge

Sapir, E. (1921). An introduction to the study of speech. *Language*, *1*, 15.

Sapir, E. (1929). The status of linguistics as science. *Reprinted in The Selected Writings of Edward Sapir in Language, Culture, and Personality/University of California P*.

Sapir, E., & Swadesh, M. (1946). American Indian grammatical categories. *Word*, *2*(2), 103–112.

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences*, *107*(6), 2408–2412.

Schaffer, W. M. (1974). Optimal reproductive effort in fluctuating environments. *The American Naturalist*, *108*(964), 783–790.

Scharfstein, L. A., Beidel, D. C., Sims, V. K., & Rendon Finnell, L. (2011). Social skills deficits and vocal characteristics of children with social phobia or Asperger's disorder: A comparative study. *Journal of Abnormal Child Psychology*, *39*, 865–875.

Schick, B., De Villiers, P., De Villiers, J., & Hoffmeister, R. (2007). Language and Theory of Mind: A Study of Deaf Children. *Child Development*, *78*(2), 376–396. https://doi.org/10.1111/j.1467-8624.2007.01004.x

Schieffelin, B. B. (2008). Speaking only your own mind: Reflections on talk, gossip and intentionality in Bosavi (PNG). *Anthropological Quarterly*, *81*(2), 431–441.

Scholz, B. C., Pelletier, F. J., Pullum, G. K., & Nefdt, R. (2024). Philosophy of Linguistics. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2024). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2024/entries/linguistics/

Schulz, J. F., Bahrami-Rad, D., Beauchamp, J. P., & Henrich, J. (2019). The Church, intensive kinship, and global psychological variation. *Science*, *366*(6466).

Schwanenflugel, P. J., Fabricius, W. V., Noyes, C. R., Bigler, K. D., & Alexander, J. M. (1994). The Organization of Mental Verbs and Folk Theories of Knowing. *Journal of Memory and Language*, *33*(3), 376–395. https://doi.org/10.1006/jmla.1994.1018

Scott-Phillips, T. C. (2010). The evolution of relevance. *Cognitive Science*, *34*(4), 583–601.

Scott-Phillips, T. C. (2014). *Speaking Our Minds: Why human communication is different, and how language evolved to make it special*. Macmillan International Higher Education.

Scott-Phillips, T. C., Kirby, S., & Ritchie, G. R. (2009). Signalling signalhood and the emergence of communication. *Cognition*, *113*(2), 226–233.

Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind eyes: An absence of spontaneous theory of mind in Asperger syndrome. *Science*, *325*(5942), 883–885.

Senzaki, S., Masuda, T., & Ishii, K. (2014). When Is Perception Top-Down and When Is It Not? Culture, Narrative, and Attention. *Cognitive Science*, *38*(7), 1493–1506. https://doi.org/10.1111/cogs.12118

Seuren, P. A. M. (1998). *Western linguistics: An historical introduction*. Blackwell Publishers.

Seyfarth, R. M., & Cheney, D. L. (2014). The evolution of language from social cognition. *Current Opinion in Neurobiology*, *28*, 5–9.

Shatz, M., Wellman, H. M., & Silber, S. (1983). The acquisition of mental verbs: A systematic investigation of the first reference to mental state. *Cognition*, *14*(3), 301–321. https://doi.org/10.1016/0010-0277(83)90008-2

Sherman, R. A., Rauthmann, J. F., Brown, N. A., Serfass, D. G., & Jones, A. B. (2015). The independent effects of personality and situations on real-time expressions of behavior and emotion. *Journal of Personality and Social Psychology*, *109*(5), 872.

Simion, F., Regolin, L., & Bulf, H. (2008). A predisposition for biological motion in the newborn baby. *Proceedings of the National Academy of Sciences*, *105*(2), 809–813.

Slaughter, V., & Perez-Zapata, D. (2014). Cultural variations in the development of mind reading. *Child Development Perspectives*, *8*(4), 237–241.

Smock, P. J., & Schwartz, C. R. (2020). The demography of families: A review of patterns and change. *Journal of Marriage and Family*, *82*(1), 9–34.

Sonneville, L. M. J. D., Verschoor, C. A., Njiokiktjien, C., Veld, V. O. het, Toorenaar, N., & Vranken, M. (2002). Facial Identity and Facial Emotions: Speed, Accuracy, and Processing Strategies in Children and Adults. *Journal of Clinical and Experimental Neuropsychology*, *24*(2), 200–213. https://doi.org/10.1076/jcen.24.2.200.989

Southgate, V., Chevallier, C., & Csibra, G. (2009). Sensitivity to communicative relevance tells young children what to imitate. *Developmental Science*, *12*(6), 1013–1019.

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, *18*(7), 587–592.

Sperber, D. (1996). Explaining culture: A naturalistic approach. *Cambridge, MA: Cambridge*.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic Vigilance. *Mind & Language*, *25*(4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x

Sperber, D., & Wilson, D. (2001). *Relevance: Communication and cognition* (2nd ed). Blackwell Publishers.

Sperber, D., & Wilson, D. (2002). Pragmatics, modularity and mind-reading. *Mind & Language*, *17*(1–2), 3–23.

Sterelny, K. (1990). *The representational theory of mind: An introduction.* Basil Blackwell.

Stewart, S. L., Schepman, A., Haigh, M., McHugh, R., & Stewart, A. J. (2019). Affective theory of mind inferences contextually influence the recognition of emotional facial expressions. *Cognition and Emotion*, *33*(2), 272–287.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K.-E., & others. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, pnas-0903616106.

Stivers, T., Mondada, L., & Steensig, J. (2011). *The Morality of Knowledge in Conversation* (Vol. 29). Cambridge Univ Pr.

Storry, M., Childs, P., & others. (2002). *British cultural identities*. Routledge London.

Sullivan, K., Zaitchik, D., & Tager-Flusberg, H. (1994). Preschoolers can attribute second-order beliefs. *Developmental Psychology*, *30*(3), 395.

Su-Russell, C., & Sanner, C. (2023). Chinese childbearing decision-making in mainland China in the post-one-child-policy era. *Family Process*, *62*(1), 302–318.

Sutrop, U. (2001). List Task and a Cognitive Salience Index. *Field Methods*, *13*(3), 263–276. https://doi.org/10.1177/1525822X0101300303

Tajima, Y., & Duffield, N. (2012). Linguistic versus cultural relativity: On Japanese-Chinese differences in picture description and recall. *Cognitive Linguistics*, *23*(4). https://doi.org/10.1515/cog-2012-0021

Tan, J., & Harris, P. L. (1991). Autistic children understand seeing and wanting. *Development and Psychopathology*, *3*(2), 163–174.

Tardif, T., & Wellman, H. M. (2000). Acquisition of mental state language in Mandarin- and Cantonese-speaking children. *Developmental Psychology*, *36*(1), 25–43. https://doi.org/10.1037/0012-1649.36.1.25

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54.

Taylor, M., & Carlson, S. M. (1997). The relation between individual differences in fantasy and theory of mind. *Child Development*, *68*(3), 436–455.

Thompson, E. C., & Juan, Z. (2006). Comparative Cultural Salience: Measures Using Free-List Data. *Field Methods*, *18*(4), 398–412. https://doi.org/10.1177/1525822X06293128

Tomasello, M. (1988). The role of joint attentional processes in early language development. *Language Sciences*, *10*(1), 69–88.

Tomasello, M. (2019). *Becoming human: A theory of ontogeny*. Harvard University Press.

Tomasello, M., & Carpenter, M. (2007). Shared intentionality. *Developmental Science*, *10*(1), 121–125.

Tomasello, M., & Farrar, M. J. (1986). Joint attention and early language. *Child Development*, 1454–1463.

Tomasello, M., Hare, B., Lehmann, H., & Call, J. (2007). Reliance on head versus eyes in the gaze following of great apes and human infants: The cooperative eye hypothesis. *Journal of Human Evolution*, *52*(3), 314–320.

Tomonaga, M., Tanaka, M., Matsuzawa, T., Myowa-Yamakoshi, M., Kosugi, D., Mizuno, Y., Okamoto, S., Yamaguchi, M. K., & Bard, K. A. (2004). Development of social cognition in infant chimpanzees (Pan troglodytes): Face recognition, smiling, gaze, and the lack of triadic interactions 1. *Japanese Psychological Research*, *46*(3), 227–235.

Tosun, S., Vaid, J., & Geraci, L. (2013). Does obligatory linguistic marking of source of evidence affect source memory? A Turkish/English investigation. *Journal of Memory and Language*, *69*(2), 121–134. https://doi.org/10.1016/j.jml.2013.03.004

Trueswell, J. C., Lin, Y., Armstrong III, B., Cartmill, E. A., Goldin-Meadow, S., & Gleitman, L. R. (2016). Perceiving referential intent: Dynamics of reference in natural parent–child interactions. *Cognition*, *148*, 117–135.

Turner, R., & Felisberti, F. M. (2017). Measuring mindreading: A review of behavioral approaches to testing cognitive and affective mental state attribution in neurologically typical adults. *Frontiers in Psychology*, *8*, 47.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Tybur, J. M., Lieberman, D., Kurzban, R., & DeScioli, P. (2013). Disgust: Evolved function and structure. *Psychological Review*, *120*(1), 65.

Udell, M. A. R., Dorey, N. R., & Wynne, C. D. L. (2011). Can your dog read your mind? Understanding the causes of canine perspective taking. *Learning & Behavior*, *39*(4), 289–302. https://doi.org/10.3758/s13420-011-0034-6

Ünal, E., & Papafragou, A. (2018). The relation between language and mental state reasoning. *Metacognitive Diversity: An Interdisciplinary Approach*, 153.

U.S. Census Bureau. (2020). *HISPANIC OR LATINO, AND NOT HISPANIC OR LATINO BY RACE*. https://data.census.gov/table/DECENNIALPL2020.P2?q=p2&g=010XX00US

Vásquez, C., & Urzúa, A. (2009). Reported speech and reported mental states in mentoring meetings: Exploring novice teacher identities. *Research on Language and Social Interaction*, *42*(1), 1–19.

Verspoor, M. H., & Pütz, M. (2000). *Explorations in linguistic relativity*.

Viberg, Å. (1984). The verbs of perception: A typological study. In B. Butterworth, B. Comrie, & Ö. Dahl (Eds.), *Explanations for language universals* (pp. 123–162). Mouton de Gruyter.

Virányi, Z., Gácsi, M., Kubinyi, E., Topál, J., Belényi, B., Ujfalussy, D., & Miklósi, Á. (2008). Comprehension of human pointing gestures in young human-reared wolves (Canis lupus) and dogs (Canis familiaris). *Animal Cognition*, *11*(3), 373–387. https://doi.org/10.1007/s10071-007-0127-y

Vonk, J., & Pitzen, J. (2017). Believing in other minds: Accurate mentalizing does not predict religiosity. *Personality and Individual Differences*, *115*, 70–76.

Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences*, *114*(43), 11374–11379.

Wellman, H. M. (2013). Universal social cognition. *Navigating the Social World: What Infants, Children, and Other Species Can Teach Us*, 69–74.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*(3), 655–684.

Wellman, H. M., & Estes, D. (1987). Children's early use of mental verbs and what they mean. *Discourse Processes*, *10*(2), 141–156.

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, *75*(2), 523–541.

Wenzel-Teuber, K. (2017). Statistics on Religions and Churches in the People's Republic of China–Update for the Year 2016. *Religions and Christianity in Today's China*, *7*(2), 26–53.

White, S., Hill, E., Happé, F., & Frith, U. (2009). Revisiting the strange stories: Revealing mentalizing impairments in autism. *Child Development*, *80*(4), 1097–1117.

Whiten, A. (1996). When does smart behaviour-reading become mind-reading? *Theories of Theories of Mind*, 277.

Whiten, A., & Byrne, R. W. (1997). *Machiavellian intelligence II: Extensions and evaluations* (Vol. 2). Cambridge University Press.

Whorf, B. L. (1956). *Language, thought, and reality: Selected writings of Benjamin Lee Whorf* (J. B. Carroll, Ed.). Cambridge University Press: MIT Press.

Wierzbicka, A. (1972). *Semantic primitives*. Athenäum.

Wierzbicka, A. (1992). *Semantics, Culture, and Cognition. Universal Human Concepts in Culture-Specific Configurations*. Oxford University Press.

Wierzbicka, A. (1996). *Semantics: Primes and universals: Primes and universals*. Oxford University Press, UK.

Wilkinson, A., Mandl, I., Bugnyar, T., & Huber, L. (2010). Gaze following in the red-footed tortoise (Geochelone carbonaria). *Animal Cognition*, *13*(5), 765–769.

Williams, B. K., Sawyer, S. C., & Wahlstrom, C. (2012). *Marriages, families, and intimate relationships*. Pearson Education.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128.

Winner, E. (1997). *The Point of Words: Children's Understanding of Metaphor and Irony*. Harvard University Press.

Wnuk, E., & Majid, A. (2014). Revisiting the limits of language: The odor lexicon of Maniq. *Cognition*, *131*(1), 125–138. https://doi.org/10.1016/j.cognition.2013.12.008

Wolff, P., & Holmes, K. J. (2011). Linguistic relativity. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(3), 253–265. https://doi.org/10.1002/wcs.104

Woo, B. M., Tan, E., Yuen, F. L., & Hamlin, J. K. (2023). Socially evaluative contexts facilitate mentalizing. *Trends in Cognitive Sciences*, *27*(1), 17–29.

Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford university press.

WSJ/NORC. (2023). *WSJ/NORC Poll March 2023*. NORC. https://www.norc.org/content/dam/norc-org/pdf2024/wsj-norc-topline-march2023.pdf

Wu, S., & Keysar, B. (2007). The effect of culture on perspective taking. *Psychological Science*, *18*(7), 600–606.

Yao, X. (2007). Religious belief and practice in Urban China 1995–2005. *Journal of Contemporary Religion*, *22*(2), 169–185.