

THE UNIVERSITY OF CHICAGO

THINKING OUT LOUD: CHILDREN USE CONVERSATIONAL CUES TO INFER
UNDERLYING MENTAL STATES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE SOCIAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF PSYCHOLOGY

BY

BEN MORRIS

CHICAGO, ILLINOIS

AUGUST 2024

© Copyright 2024 by Ben Morris

All Rights Reserved

Table of Contents

List of Figures.....	iv
Acknowledgements	v
Abstract.....	ix
Introduction.....	1
<i>Dissertation Research.....</i>	<i>9</i>
<i>Outline.....</i>	<i>10</i>
Chapter 1: Inferring others' internal states.....	14
<i>Experiment 1a.....</i>	<i>19</i>
<i>Experiment 1b.....</i>	<i>24</i>
<i>Experiment 2a.....</i>	<i>26</i>
<i>Experiment 2b.....</i>	<i>34</i>
<i>Experiment 3.....</i>	<i>40</i>
Chapter 2: Reasoning about others' expectations to learn stereotypes.....	53
<i>Experiment 1.....</i>	<i>56</i>
<i>Experiment 2.....</i>	<i>60</i>
<i>Experiment 3.....</i>	<i>66</i>
Chapter 3: Listener design and listener knowledge	75
<i>Experiment 1a.....</i>	<i>78</i>
<i>Experiment 1b.....</i>	<i>80</i>
<i>Experiment 2.....</i>	<i>85</i>
General Discussion.....	98
References	107

List of Figures

Figure 1. Example stimuli used in Experiments 1a and 1b.....	20
Figure 2. Age-binned results from Experiments 1a and 1b	22
Figure 3. Example stimuli used in Experiment 2 for the Preference trial.....	30
Figure 4. Adults' responses from Experiment 2a	32
Figure 5. Preference trial results from Experiments 2a and 2b.....	37
Figure 6. Preference condition results from Experiment 3	42
Figure 7. Labelling condition results from Experiment 3	43
Figure 8. A still from Experiment 1 showing the basic experimental setup.....	58
Figure 9. Children's toy selections across conditions for Experiment 1	59
Figure 10. A still from Experiment 2 showing the Hibbles.....	62
Figure 11. Children's weirdness judgements for Experiment 2	63
Figure 12. Children's responses for the teasing measure for Experiment 2	65
Figure 13. A schematic showing the logic for for Experiment 3	68
Figure 14. Adults' judgements for the weirdness and teasing measures Experiment 3.....	69
Figure 15. Example stimuli showing the basic trial structure layout for Experiment 1b	83
Figure 16. Age-binned results for Experiment 1b	84
Figure 17. Data for Experiment 2 showing children's judgements for the knowledge and appraisal measures	91

Acknowledgements

I've spent the last six years of my professional life asking questions about a few small things. In this dissertation, I examine how we come to see the minds that lie behind the words others use. Like my work itself, the realities of this dissertation largely lie unsaid behind the words that will follow. Here, I do my best to offer a glimpse behind those words.

This dissertation is indebted to former mentors who got me started on this path. I am grateful to Kathryn Oleson, Daniel Reisberg, Michael Frank, Molly Lewis, and Claire Hughes. I am especially grateful to Jennifer Henderlong Corpus who introduced me to questions I didn't know could be asked. I would also like to thank my former advisor, Dan Yurovsky. I wouldn't be where I am if it weren't for you. Working with you convinced me to go to graduate school and continues to shape the kind of science I do.

When I first started at UChicago, it was Dan, me, and some empty offices. Thank you to the members of CaLLab, as it was (Claire Bergey, Ashley Leung, Xiuyuan Flora Zhang, Jo Denby), for bringing those offices to life. With you, I first really saw what doing science could look like. To my labmates in the DIBS lab (Hannah Kim, Mitchell Landers, Katie Vasquez, and Alex Mackiel), this work would not have been the same without you. To our wonderful lab managers over the years (Molly Gibian, Gemma Smith, and Isabella Ramkissoon) and the many research assistants who came on this journey with me (most especially Cassie Wilson and Peter Pan), this work wouldn't have been possible without you.

I am also deeply grateful to the broader community I have found in our department. To Casey, Carol, Claire, Hannah, Izzy, Jenny, Jimmy, Jo, Kaila, Letty, Rachel, Rad, Yağmur and many others—thank you for your friendships that have brought so much warmth to our florescent halls. To those I've looked up to over the years—especially Kensy, Michelle, Ruthe,

and Will— thank you for showing me the way. To Ray, Smiley, and Kristi—thank you for keeping our department afloat.

This dissertation would not have been possible without large quantities of coffee and romantic working atmospheres to while away the hours. Thank you to everyone who made me a cup of coffee—most especially at Grounds of Being, Plein Air, Dollop (State/Van Buren), Intelligentsia (Logan Square), Gaslight, and Heritage Bikes. And thank you to many beautiful spaces that have shepherded me and this work, including the aforementioned coffeeshops, Mansueto, Harper, and the ISAC. Most of this dissertation was written in one of those places.

I am grateful to my wonderful committee—Amanda Woodward, Susan Goldin-Meadow, and Marisa Casillas. Amanda, every time I've talked to you about this work you have offered truly incisive comments that I'll be mulling over for years to come. Susan, yours is the lab where I really saw what a large, engaged scientific community could look like. I'm continually astonished by your ability to keep asking the big, hard questions, and being as interested in the questions as the answers. Middy, it would be hard to overestimate the impact you have had on me. You've taught my favorite classes, gotten me to read some of my favorite papers, and pushed me to have think about my work in new ways. Your arrival at UChicago was almost perfectly timed, and you've become a dear and important mentor to me.

And to my advisor, Alex Shaw, it is hard to capture how grateful I am to be your student. From the moment we started working together, I never doubted that you were in my corner, rooting for me and genuinely looking out for me. After six years and countless meetings together, I see your influence everywhere—in the questions I ask, the way I make a talk, and the critiques I generate. I am confident that I will be noticing mannerisms (big and small) that I've inherited from you for years to come. I am grateful for the freedom you gave me to ask my own

questions, your faith in my ability to pursue them, and your enthusiasm to think through them with me. Thinking out loud with you will always be one of my favorite things.

Thank you also to my family, for their love and support:

To my siblings, who I am lucky to call friends. Julian, my big brother, thank you for looking out for me and looking for me. Meghan, my big sister, thank you for showing me the way and for being there to listen when I get lost.

To Ryan, I'll never know how you wrote one of these with a baby and a pandemic.

To Finn, River, & Genevieve, for being new lights. I can't wait to keep knowing you.

To my parents, for all you've done for me and all you wish for me. Mom, thank you for giving me my red beard and my sweet tooth, for being the one I call when I need a pep talk, and for loving freely and generously. Dad, thank you for giving me a love of words and pretzel rods, for being my human dictionary, and for showing me how to exude quiet warmth and thoughtfulness.

And thank you to my friends, who have seen me through (including many beyond this list):

To Julia Aizuss, for your camaraderie and for the way you find your words.

To Scott Daniel, for your radiance and for your seeing radiance in those around you.

To Cameron Knox Day, for your easy friendship and your snort-filled laugh.

To Kaylie Engel, for swooping into my life and brightening every corner.

To Casey Ferrara, for complaining with me and making it better.

To Michael Galperin, for sitting next to me in Mansueto and making me smile.

To Ashley Leung, for being my model of finding balance.

To Dan Medvedev, for your indifference toward what doesn't matter and for breaking through to me.

To Tess Tumarkin, for calling me and for being the one I know will always swim.

To Thea Rice, for really hearing me and always remembering. And for always going to the movies with me. Our friendship is now written so deeply into my life that it's sometimes hard to remember a beforetime. In all the smallest moments, being with you reminds me how to listen, how to laugh, how to care, and how to be good.

To Yağmur Deniz Kısa, for bringing about my silliest laughs and my deepest calms, by turns and in equal measure. Thank you for all the chitchats, big and small, and for reminding me how much I love to play. Thank you also for being the first to leave—I miss you often but seeing you leaping always gives me courage to step.

To Claire Bergey, for giving me new things to look at, for giving me new ways to see them, and for sitting next to me while we look. Yours is the friendship that pulled me in first in Chicago; your magnetism and self-evident style give you a gravitational pull. It was talking with you that taught me just how fun our work could be. Thank you for your thoughtfulness, your inquisitiveness, your laughter, and your care. My world opened wider through being your friend.

To Hannah Kim, for saying yes when I'm still wavering. The marks of our friendship are many—in my gestures, in my passport, and on my hair. I think I could have done this without you, but I can't imagine it and it would have sucked. When I'm with you I laugh more, second-guess myself less, dance harder, and relax more deeply. I have become more myself by knowing you.

I love you all. Thank you for helping me think, helping me write, and helping me relax.

Abstract

In conversation, much is communicated without being directly said. By leveraging an understanding of how language relates to mental states and processes, communication becomes a window into a speaker's thinking. In this dissertation, I demonstrate how children (ages 4-to-9) come to readily reason about others' mental states based not just on what they say, but how they say it—from how easily something is said (Chapter 1), to how surprised someone seems (Chapter 2), and even how someone is spoken *to* (Chapter 3). In Chapter 1, I explore the humble “um.” While disfluencies in speech are often overlooked as meaningless errors by laypeople and researchers alike, I demonstrate that children interpret disfluencies as socially meaningful—over and above the content of what is said—and use them to flexibly infer a speaker's knowledge and preferences. In Chapter 2, I explore how children reason about the implications of conversational cues in *feedback*, specifically how markers of surprisal and production difficulty (e.g., “Oh! Um... Sure”) lead children and adults to infer a speaker's underlying expectations. I find that conversational feedback not only signals a speaker's expectations, but also provides an unappreciated avenue for the transmission of social beliefs and stereotypes. In Chapter 3, I show that how someone is spoken *to* may shape the mental inferences that children make about that person before that person ever says a word. When a speaker offers basic categorical information, children and adults infer that the *listener* is likely unfamiliar with the topic at hand. Across these three chapters, I argue that children are actively, rationally, and flexibly inferring mental states by integrating subtle conversational cues, context, and prior discourse. Capitalizing on their skills as budding mentalists, children are learning to extract social meaning from subtle conversational cues—skills that are fundamental to becoming smooth conversationalists and sophisticated social learners.

Introduction

Conversation is perhaps the most ubiquitous social behavior of human life. While language itself has been a fundamental area of study since the founding of psychology, theoretical accounts and empirical studies of language have often ignored its most simultaneously mundane and complex form: social conversation. Indeed, foundational accounts of language largely treated language as a formal system devoid of or at least separable from social context (e.g., Chomsky, 1959). On the other hand, pragmatic accounts of language have shifted this narrative to emphasize the importance of considering language *use*, conversational context, and social reasoning (Clark, 2009; Grice, 1975; Goodman & Frank, 2016; Sperber & Wilson, 1995). These accounts suggest that even relatively simple uses of language (e.g., referential expressions) go well beyond literal, semantic meaning by recruiting social-inferential reasoning (Frank & Goodman, 2012). Under such accounts, children’s language use, understanding, and learning are powerfully shaped by their social reasoning skills—skills that are evident even in the second half of the first year of life (Bohn & Frank, 2019; Tomasello, 2008). In this dissertation, I probe this important connection between communication and social reasoning in a different way; rather than exploring how children leverage social reasoning and mental states to learn language and converse with others, I explore how children leverage subtle conversational cues to learn about people and their mental states.

Communication and Mental States

Early communication seems to be fundamentally mentalistic, as exemplified by two literatures with infants even before they begin to speak themselves. The first focuses on how language activates infant’s *own* mental representations and argues infants learn to map words to associated mental concepts, rather than just to concrete objects. For example, young infants are

able to communicate about absent entities (e.g., Ganea & Saylor, 2013), and understand something of the intentions behind communicative acts (e.g., Behne, Carpenter, & Tomasello, 2005). The second line of research focuses on monitoring *others'* mental representations in the service of communication, based on tasks that manipulate an interlocutor's knowledge state. For example, infants gesture more for a toy when their interlocutor is ignorant about a hidden toy's location, compared with when their interlocutor also saw it being hidden (Liszkowski et al., 2007; Liszkowski et al., 2008; O'Neill, 1996). From its earliest developmental roots, communication seems to be grounded in social reasoning, and these skills are argued to form the basis for pragmatic development (Bohn & Frank, 2019).

Broadly, research in pragmatics focuses on how social reasoning allows speakers and listeners to go beyond literal meaning, relying on context and inference rather than just the literal semantics of what someone says. As we have already seen, the beginnings of pragmatic reasoning are already evident in infancy, with infants integrating context and other's knowledge into their own communication (Liszkowski et al., 2008; O'Neill, 1996). Infants also demonstrate a nascent understanding of common ground, tracking shared experiences and reasoning about the implications of shared (vs. private) knowledge. For example, when 24-month-olds hear a novel word, they expect the speaker to be referring to a new object, and that newness depends on the speaker's prior experience rather than the child's own (Akhtar et al., 1996). Even before language gets off the ground, social reasoning seems to already subserve children's communication and understanding.

While research with infants suggests that by 12-months they are engaging with language through mental state reasoning, research in experimental pragmatics with older children suggests they continue to struggle with more complex tasks (e.g., implicature, metaphor, hyperbole) well

into the preschool years and beyond (for review spanning this age range see Bohn & Frank, 2019). Tasks of pragmatic implicature require children to derive an implied meaning beyond the literal content of what someone says. For example, when someone says they “ate some of the cookies”, adults infer that they did not eat all of the cookies (e.g., scalar implicature Noveck, 2001) and when someone refers to a dax as a “large dax”, adults infer that daxes are usually smaller (e.g., Bergey & Yurovsky, 2023). However, young children seem to struggle with these kinds of implicatures and are more willing to endorse literal interpretations than adults (e.g., Noveck, 2001). These kinds of implicatures are thought to rely on social reasoning—interpreting what someone said via a process of recursive reasoning that integrates the context at hand, relevant alternative things they could have said, and their desire to be understood (as formalized by the Rational Speech Act framework; Goodman & Frank, 2016). Indeed, pragmatic implicatures seem to critically intersect with reasoning about the speaker’s knowledge, and even children as young as 4-to-5 seem to understand this in tasks that clarify the alternatives (e.g., Papafragou et al., 2020). Scholars have argued that pragmatics represents a special application of social cognition—i.e. recruiting our more general mental state reasoning skills—and even argued for its continuity across development (e.g., Bohn & Frank, 2019).

Research into pragmatics makes it clear how reasoning about others’ mental states is inextricably linked to how we communicate and what we communicate about; even infants’ earliest communications seem to recruit mental representations of others. However, these literatures provide a very limited evidence base for understanding *how* and how *flexibly* infants and children infer others’ mental information. The extant literature largely operationalizes and manipulates knowledge in similar ways, appealing to a kind of seeing-as-knowing rule (e.g., someone is ignorant because they failed to witness a critical event). For the current purposes, it is

especially important to emphasize that these studies rarely use *language* cues to instantiate or manipulate mental state information (c.f. Koenig & Echols, 2003). As a result, these studies cannot address how conversational cues themselves license mental state inference. In other words, this literature makes clear that epistemic reasoning is involved *in* early communication; however, we still know surprisingly little about how epistemic reasoning *from* communication happens in development.

A series of studies from Vouloumanos and colleagues provides evidence of the earliest roots of children's ability to extract mental information from communication itself. In these studies, infants by at least 12 months expect that language can transmit information and intentions, and expect a listener to act accordingly (Martin et al., 2012; Vouloumanos et al., 2012). Infants generalize these expectations to a variety of communicative acts (e.g., speaking, pointing), and understand that non-communicative behaviors do not have these same capacities (e.g., coughing; Martin et al., 2012; Krehm et al., 2014). Infants even understand some critical constraints on communication— that information flows from speaker to listener (Martin et al., 2012), that it requires perceptual access (e.g., someone has to see you pointing; Krehm, et al., 2014), and that speakers must share a language (Pitts et al., 2015; Colomer & Sebastian-Galles, 2020).

Across these studies, preverbal infants perhaps even as young as 6 months (Vouloumanos et al., 2014) demonstrate a set of expectations for how acts of communication relate to mental states. These skills may thus serve as precursors to richer, flexible conversational mental reasoning later in development. That is, the ability to extract complex mentalistic information from communication is likely underlied by these simpler, nascent understandings of the relationship between language and mental states that we see in infancy. We next turn to review

the literatures exploring children’s abilities to derive these more complex inferences from language. How do children come to understand that what we say (and how we say it) reflects how we’re thinking?

Trust in Testimony

By the preschool years, children are clearly able to extract more nuanced mentalistic information from language itself. That is, while infants reason about the presence or absence of a vocalization, young children evaluate the *content* of a claim itself and its relation to the speaker’s underlying knowledge state. The vast literature on trust in testimony highlights this emerging understanding of how language may convey mental states, especially knowledge (for review see Sobel & Kushnir, 2013). In the dominant paradigm, children are presented with speakers who label a series of familiar items with either accurate, conventional labels or incorrect, non-conventional labels (e.g., referring to a ball as a “shoe”). Across a wide variety of studies, children by age 3-to-4 preferentially seek new information from previously accurate speakers and, when presented with contrasting information, selectively endorse information from previously accurate speakers (e.g., Koenig & Harris, 2005). While the beginnings of this reasoning may even be evident in infancy (Henderson et al. 2015; Koenig & Echols 2003; for a review across ages see Harris et al., 2018), children by middle childhood are able to reason about knowledge based on subtle aspects of someone’s claim—such as whether it conveys generalizable information and whether it is based on mere observation (Aboody, 2022; Cimpian & Scott, 2012; Koenig et al., 2015).

A related literature that examines children’s burgeoning understanding of the relationship between language and mental states focuses on children’s sensitivity to *explicit* markers of epistemic states, such as evidential claims and mental-state verbs. For example, when an agent’s

mental states are communicated explicitly through their word choice (e.g., “This is a spoon” vs. “I think this is a spoon”), children by age 3 are less likely to trust the claim of the uncertain speaker (e.g., Jaswal & Malone, 2007; Sabbagh & Baldwin, 2001). Children also discount speakers who profess ignorance about familiar items, e.g., saying “I don’t know” (Sabbagh & Shafman 2009). Importantly, children distinguish between ignorance claims and inaccurate claims, being willing to endorse new information from previously ignorant speakers, but not from previously inaccurate speakers (Kushnir & Koenig, 2017). In this way, children seem able to monitor explicit knowledge claims and their implications.

While speakers may sometimes directly comment on their mental states, these situations likely account for a small slice of the instances of epistemic inferences that are routine in conversation. Instead, *how* something is said may often be as crucial as what is said. Rather than explicitly marked mental signals, much of conversational inference is more likely rooted in these subtler, paralinguistic signals that exist “around the edge of language” (Bolinger, 1964). Research on evaluations of speaker accent provides one such case of a paralinguistic cue in communication that seems to license epistemic inferences from early in development. Infants show robust social preferences for the native-accented speakers over foreign-accented speakers (Kinzler et al., 2007), preschool age children endorse information from more from native-accented speakers (Kinzler et al., 2011), and older children even show complex accent-based stereotypes (such as “Northern = smart, Southern = nice” for American English; Kinzler & DeJesus, 2012). Accent provides relevant but broad, informant-level information; however, in most of conversation we need to engage in specific mental state reasoning based on communicative cues in-the-moment.

Another important body of work that takes a more in-the-moment approach to study paralinguistic interactional cues focuses on confidence displays. While young children don't seem to understand the implications of explicit mental state expressions of confidence (e.g., "I think" vs. "know") until about 4 years of age (Jaswal & Malone, 2007), research focused on broader, paralinguistic cues of confidence demonstrates that children as young as 24 months are already sensitive to an agent's confidence (Birch, Akmal, & Frampton, 2010; Brosseau-Liard & Poulin-Dubois, 2014). These studies exploit multimodal cues to uncertainty—shrugging gestures, scrunched facial expression, behavioral hesitation, and more cues. By age 2, children selectively imitate demonstrations from a confident agent, rather than an agent who displays uncertainty (Birch et al., 2010; Brosseau-Liard & Poulin-Dubois, 2014).

While the research on trust in testimony has been highly informative, most of this work both the explicit and less explicit, takes an informant-centered approach, wherein the key questions concern young children's skepticism and trust of a particular speaker. Studies commonly manipulate prior accuracy or manipulate trait-like qualities (speaker race, familiarity, age, niceness and more), and look at subsequent behaviors that signal trust. Instead, I advocate for an online approach, wherein the key questions concern how children evaluate a particular utterance in-the-moment from a given speaker. In other words, we should be less interested in the factors that lead children to make broad inferences about a speaker, and more interested in the factors that lead them to make inferences about the mental states behind what that speaker is saying now.

The need for a more online approach is especially important in thinking about development, as evident when you consider that very young children typically interact with a restricted number of people. Early problems of social reasoning from conversation likely have

less to do with forming an evaluation of the speakers around you based on prior behavior or static traits, and more to do with learning to monitor what they're saying now and what that indicates based on in-the-moment cues. When a mother tells her young child where to find his shoes, the question isn't how trustworthy she is but instead how confident her claim seems in this moment—does she sound like she definitely knows or is she merely guessing? Real world informants are not entirely trustworthy or untrustworthy, but instead what they say may be trustworthy in certain moments, but not others. While past work acknowledges and moves beyond these constraints (e.g., Jaswal & Malone, 2007; Kushnir et al., 2015; Kushnir & Koenig, 2017; Liberman & Shaw, 2020), the core approach across this area remains informant-centered.

Language is famed for its unique capacity to transmit information between agents, and thus what someone says (and especially *how* they say it) can reveal a great deal about their internal world. The starts, stops, and idiosyncrasies of how we speak can reflect our mental processes leaking out in real time. Lay people seem to intuitively understand some of the systematicity of this relationship between the mental processes and language; adult listeners use aspects of how someone speaks to reason about a speaker's ongoing mental processing (e.g., whether they are uncertain, recalling something, fabricating an answer, or planning how to say something with tact; Barr, 2003; Brennan & Williams, 1995; Clark, 1996; Fox Tree, 2002; Roberts et al., 2011; Ziano & Wang, 2021). We know a lot about the kind of mental state reasoning that allows a young child to communicate, but we know surprisingly little about how young children can extract this rich information about others' minds based on communicative signals alone.

Dissertation Research

Across multiple literatures, it is clear that children are beginning to understand the relationship between language and mental state reasoning—understanding that speech transmits information as infants (e.g., Martin et al., 2012), flexibly imitating on the basis of confidence as toddlers (e.g., Birch et al., 2010), and even using the form of a claim to evaluate a speaker’s knowledge as young children (e.g., Koenig et al., 2015). Together, these findings present the intriguing possibility that children are developing a rich, flexible model of the relationship between a speaker’s internal mental states and their utterance. Such a model should not only allow children to incorporate mental state information in the service of communication, but also critically to generate inferences from what a speaker says (and how they say it) back to the internal representation that generated that utterance.

In this dissertation, I explore how children come to reason about what someone says as reflective of their thinking, using the subtleties of how something is said to make inferences about a person’s underlying mental state at that moment. As has long been noted by researchers in theory of mind, mental states are not directly observable, but must be inferred from observable action. Utterances are a unique class of actions that are particularly rich with real-time mental state information—not just in explicit markers like mental state verbs, but also in the complex tapestry of subtle, paralinguistic cues through which our inner lives spill out (Barr, 2003). Across three chapters, I demonstrate how children begin making mental inferences *from* communicative cues directly. I argue that children’s reasoning reflects a flexible, rational inference process about how language is used by integrating context, subtle cues, and prior discourse to derive distinctive implications about a speaker’s internal states. By leveraging an understanding of how language

relates to mental states and processing, communication becomes a window into a speaker's thinking and processing.

This kind of mental state reasoning in conversation is crucial for children to become savvy conversationalists. However, this dissertation underscores how this reasoning also provides key information for early social learners—offering a glimpse into speaker's knowledge, preferences, expectations, and appraisals of others. In three chapters, I demonstrate children's (ages 4-9) burgeoning ability to draw social inferences from largely paralinguistic cues—about how easily something is said (Chapter 1), how someone responds (Chapter 2), and even how someone is spoken *to* (Chapter 3). Conversation is rife with social meaning, and this work explores how children begin to extract that meaning from subtle conversational cues.

Outline

In Chapter 1, we explore the humble “um.” While pauses and disfluencies in speech are often overlooked as meaningless errors by laypeople and researchers alike, these subtle cues are profoundly structured and rich with social meaning. Past work with children has explored how disfluencies may prompt language inferences (e.g., reference resolution, Kidd, White, & Aslin, 2011), but our work asks instead how children use these cues to reason about a speaker's mental states. Across 3 experiments (total n = 305 4- to 9-year-olds), we demonstrate that children as young as 4-5 are already interpreting disfluent pauses as socially meaningful, over and above the content of what was said. These disfluencies reflect a speaker's general, ongoing mental processes, and thus children can use them to reason about various mental states-- both speaker knowledge and preference. Overall, we see that children at all tested ages demonstrate some ability to reason contextually about the meaning of a disfluency, with older children drawing remarkably flexible inferences.

In Chapter 2, we explore how children reason about the implications of conversational cues in *feedback*. Specifically, we ask how children detect and reason about linguistic expressions of surprisal by contrasting permissive feedback statements that differ in the presence of interjections and disfluencies as two markers of surprisal (e.g., “[Oh! Um...] Sure honey”). In Experiment 1 (n = 120, 4- to 9-year-olds), we look at gender stereotypes as a real world test-case, and demonstrate that children by age 6-7 use others’ surprise to reason about others’ gender stereotyped expectations. In Experiment 2, we demonstrate that surprisal cues could serve as a mechanism for children (n = 120, 4- to 9-year-olds) and adults (n = 80) to learn expectations in the first place about novel alien groups. In Experiment 3, we combine these findings to demonstrate that adults (n = 150) readily learn novel gendered stereotypes by tracking others’ surprisal reactions. Across these experiments, we see converging evidence that conversational feedback may provide a crucial and unappreciated avenue for the transmission of social beliefs.

In Chapter 3, we ask how conversational cues may shape social inferences even before someone speaks—specifically by reasoning about how someone is spoken *to*. Studies on the development of listener design demonstrate that children by age 4-5 tailor the information they provide differently for a knowledgeable versus an ignorant listener (e.g., Baer & Friedman, 2018). In this chapter, we explore how children make the inverse inference to reason about a listener’s knowledge based on what is said to them. In Experiments 1a and 1b, adults (n = 60) and children by age 6-7 (n = 60 4- to 9-year-olds) infer that a listener who is told basic information about a familiar toy is less knowledgeable than someone who is told a non-basic explanation. In Experiment 2, we tested these inferences in a new paradigm that also tests the extent to which children see these utterances as reflective of the speaker’s *belief* about the listener or of the listener’s knowledge *per se*.

Together, these three chapters demonstrate how children come to understand and reason about the mental states and processes behind what someone says. Each chapter takes one fundamental principle of how language is used—principles we know even young children are engaging with in their own speech—and asks how they understand and invert those principles as listeners to reason about the mental processes behind utterances. First, conversation is fundamentally structured by an expectation of timeliness (Stivers et al., 2009). Developmentally, preverbal infants show rapid turn-taking patterns in their own vocalizations (e.g., Hilbrink et al., 2015), and young children are able to make rapid, anticipatory predictions about turn-taking (Casillas & Frank, 2017). In other words, young children seem to know that language has to happen fast. Chapter 1 asks how children exploit this understanding to reason about what it can mean when language comes slowly and stilted, with markers of disfluency. Second, conversation is fundamentally predictive, with speakers and listeners making rapid, online plans and predictions about what will be said next (e.g., Kuperberg & Jaeger, 2020). Even infants seem to show some signatures of predictive processing, for example, using prior sentence context to anticipate what comes next (for review across development see Zettersten, 2019). In other words, young children seem to know that conversation involves thinking about what happens next. Chapter 2 asks how children exploit this understanding to reason about what it can mean when someone’s reaction evinces unexpectedness. Third, conversation is fundamentally calibrated—what we say (and what we don’t) depends on what our listener already knows (and what they don’t; Clark & Murphy, 1982). Already by 12 months infants give more information when someone is ignorant (Liszkowski et al., 2007), and by 4-to-5 children offer different kinds of information based on what the other person knows (Baer & Friedman, 2018). In other words, young children seem to know that conversation involves accounting for what your interlocutor

knows. Chapter 3 asks how children exploit this understanding to reason about what someone knows based only on how they're spoken *to*.

Chapter 1: Inferring others' internal states

Imagine you ask someone on the street, “Where is the nearest train station?” and they reply “It’s... um... that way.” While plausibly correct, the stranger’s reply is slow and marked, indicating uncertainty about the accuracy of their statement. Despite getting a relevant answer, you yourself may still feel uncertain, perhaps choosing to confirm the direction by consulting another stranger or checking your smartphone. In this way, adults can use disfluencies to make inferences about the inner workings of a speaker’s mind, which is an important social cognitive skill that can shape how they learn from and evaluate others. *How* (and how quickly) something is said may often be as meaningful as *what* is said. When in development can children use disfluencies to make inferences about others’ mental states? In Chapter 1, we explore this question in three studies with children ages 4- to 9-years old.

In the example above, the stranger seems uncertain because they spoke disfluently, pausing in the middle of their utterance. When it comes to how something should be said, conversation is profoundly structured by an expectation of timeliness. Across at least 10 typologically diverse languages, conversation is marked by consistent, brief silences between speakers’ contributions—just 200ms on average (Stivers et al., 2009). However, producing language in a timely manner often conflicts with demands on processing, planning, and many other factors liable to cause delays. When such delays are unavoidable, speakers frequently produce disfluencies (e.g., filled pauses “uh” and “um” in English, among other types of disfluencies). Conversation is rife with these small disruptions; some estimates suggest that 6 disfluencies occur every 100 words in adult conversations (Fox Tree, 1995; Shriberg, 1996). For the speaker, disfluencies can reflect slowed lexical retrieval, ongoing utterance planning, and a desire to hold the conversational floor, among other proposed functions (Clark & Fox Tree,

2002; Smith & Clark, 1993). For the listener, disfluencies heighten lexical processing (Fox Tree, 2001) and are used to predict upcoming referents—e.g., ones that are generally unfamiliar or new to the conversation at hand (Arnold, Fagnano, & Tanenhaus, 2003). Given the connection between disfluencies and a speaker’s mental processes, disfluencies can act as a powerful cue that the listener can use to make inferences about the cognitive process that generated them.

From an early age, children detect and produce disfluencies, and reason about them to make language predictions. By 22 months, infants distinguish between fluent and disfluent speech, suggesting that children are already sensitive to disfluency in the second year of life (Soderstrom & Morgan, 2007). Young children also have first-hand experience with disfluencies in their own productions—prevalent from at least age 2 (Casillas, 2014). Still more impressive, by their second birthday, toddlers are able to monitor these cues online and predict that a disfluent speaker will likely refer to a novel object that is new to the discourse, rather than a familiar one that has already been discussed (Kidd et al., 2011). In the preschool years, these predictions become increasingly sophisticated— with young children blocking disfluency-based language prediction appropriately for forgetful speakers (Orena & White, 2015) or distracted speakers (Yoon & Fisher, 2020), and even drawing speaker-specific predictions (Yoon, Jin, Brown-Schmidt, & Fisher, 2021). In word learning, children as young as 3 use disfluency to guide selective novel word learning from two contrasted speakers (White, Nilsen, Deglint, & Silva, 2020). Young children seem to understand something of the relationship between disfluencies and a speaker’s mental states, at least when it comes to processing language.

While disfluencies are clearly powerful cues for predicting language, disfluencies reflect general processing delays and thus can also underwrite broad inferences about other mental processes. Indeed, adults sometimes interpret disfluencies in responding as indicative of less

comfort with the topic at hand, or even less honesty (Fox Tree, 2002; see also Ziano & Wang, 2021). Further related work examining silent pauses comes from a conversation analysis approach (Sacks, Schegloff, and Jefferson, 1974). When responding to requests (e.g., “Can you give me a ride?”), a speaker who pauses before accepting is seen as less willing to accept the request (Roberts, Francis, & Morgan, 2006; Roberts, Margutti, & Takano, 2011). Similarly, when agreeing with another’s past statement (e.g., “The flyers look good.”), delays in responding are seen as indicating less agreement (Roberts et al., 2011). These effects generalize across multiple languages and speaker judgments are even titrated by the degree of delay– the longer the delay, the less willing a speaker seems (Roberts et al., 2011).

Adults also use disfluencies, including filled pauses (e.g., “um”), to infer a speaker’s likely knowledge. When answering a factual question (e.g., “What is 5 times 7?”), disfluencies indicate a delay in searching for an answer (“Um... 35”). As speakers, adults who produce slower, disfluent – but still accurate – responses to factual questions report less knowledge about the topic at hand (Smith & Clark, 1993). As listeners, adults pick up on these kinds of cues and infer that a speaker is less knowledgeable even though the content is accurate (Brennan & Williams, 1995). Of course, adults do not view disfluencies as a simple heuristic that always signals a lack of knowledge. Many disfluencies in naturalistic speech are related to other factors, such as discourse history, speech rate, or interlocutor familiarity and may not indicate knowledgeability (Clark & Fox Tree, 2002; Shriberg, 1996). Indeed, even in responses to factual questions, adults do not always infer that disfluencies indicate incompetence, for example when answering a question with a non-answer (e.g., “Uh... I don’t know”). In this context, adults seem to even treat disfluency as a signal of *greater* knowledge– the hesitation implied by a disfluency suggests the speaker feels like they might know the answer and attempted to search for it

(Brennan & Williams, 1995; Smith & Clark, 1993). In sum, these results suggest adults may be reasoning about the underlying mental process that delays speech in a powerfully flexible, inferential manner. Holding constant what a speaker says, the speed and fluency with which a speaker says it licenses a range of contextualized inferences about their mental states.

Can children use others' pauses to form these nuanced inferences about others' mental processes? While no work has directly examined the social inferences that children generate from speech disfluencies (e.g. how someone gives a correct answer), the large literature on trust in testimony demonstrates that young children judge others knowledge based on the accuracy of *what* they say (e.g., Koenig & Harris, 2005; see Sobel & Kushnir, 2013 for a review). However, prior accuracy information may often be unavailable to the developing learner, or they may have limited knowledge on which to assess accuracy. Moreover, in many situations, a speaker's accuracy alone may convey little about their underlying knowledge of the given domain (e.g., when answering an easy question). Here, a better cue might be the speaker's ease and fluency in discussing the topic at hand. That is, children may make inferences based not just on what a person says, but on *how* (and how quickly) someone says it.

Research from the testimony literature makes it clear that children consider cues beyond accuracy when judging speaker knowledge or figuring out what to learn. Children use informant trait information, such as niceness or informant age, to guide their selective learning (e.g., Lane, Wellman, & Gelman, 2013; VanderBorghet & Jaswal, 2008). Relatedly, there is also work suggesting that children tend to believe that a native speaker is more knowledgeable than a non-native speaker (e.g. Kinzler, Corriveau, & Harris, 2011) and that people with northern accents are "smarter" than people with southern accents (Kinzler & DeJesus, 2013). Further, 2-year-old children are adept at picking up on non-verbal cues to uncertainty, preferring to imitate actions

from a confident agent rather than an agent who shrugs, hesitates, has a puzzled facial expression, and employs other cues to uncertainty (Birch, Akmal, & Frampton, 2010). Speech disfluencies present a subtle, contextualized cue to speaker knowledge and mental states more generally, thus opening a question about how children interpret the social implications of such disfluencies and whether they are taken as heuristic cues or reasoned about in an inferential manner.

An adult-like understanding of speech disfluencies requires flexible, inferential reasoning about the underlying causes of a delay based on the surrounding context. Mastering the complex social implications behind speech disfluencies can help people understand how much someone knows about the article they're telling you about, their true feelings about your new haircut, or their willingness hang out on Friday night. For young children, such skills not only enable them to become skilled conversationalists, but also to gather key information from the social world. And there are abundant opportunities for children to learn from disfluencies given their ubiquity in human conversation (Fox Tree, 1995; Shriberg, 1996). As a learner, you would need a great deal of information to learn about someone's preferences or knowledge if you could *only* observe someone's accuracy or selections. With speed and fluency information however, you can glean a great deal of information from even one instance. Disfluencies reflect ongoing mental processes, and reasoning about those processes in context yields a broad range of social inferences. When do children pick up on the notion that a simple "uh" or "um" can speak volumes about what a person knows, prefers, or believes?

In the present studies, we test how and when young children use speech disfluencies to make social inferences about speakers across a range of conversation contexts. We show that children infer an accurate, but disfluent speaker is less knowledgeable than an accurate and

fluent speaker (Experiment 1a). Further, we find that even the youngest children we tested are not relying on a heuristic rule that disfluency always implies incompetence, and instead block this inference in certain conversational contexts, such as non-answers (Experiment 1a). Children also extend this reasoning beyond the domain of knowledge; by age 6-7, children similarly infer that when someone disfluently states their preference, they may have a weaker preference (Experiments 2b and 3). We argue that children's reasoning in these tasks is fundamentally inferential and show that conversational context can prompt radically different broader inferences about the meaning of a disfluency (Experiment 3). Together, these results suggest children engage in flexible, inferential reasoning about not just what a speaker says, but the speed and fluency with which it is said.

Experiment 1a

In Experiment 1a, we ask how children ages 4- to 9-years-old use speech disfluencies to infer knowledgeability, and whether they block that inference appropriately in certain contexts—as adults do (Brennan & Williams, 1995). In all conditions, children were presented with two speakers. Speaker accuracy was held constant but we varied the response fluency and asked children which of the two speakers knows more about the topic. Children were randomly assigned to either the Labelling condition (in which the speakers accurately label an animal) or Ignorance condition (in which the speakers say “I don't know”).

Based on pilot data and previous adult work (Brennan & Williams, 1995), we predicted that children would infer an accurate but disfluent speaker would be less knowledgeable (Labelling condition). We also predicted that children would make no such inference in the Ignorance condition, and that older children may even infer that the disfluent speaker knows more in this condition, as suggested by prior adult work (Brennan & Williams, 1995). We tested

4- to 9-year-olds because there is research indicating that they clearly make complex inferences in the language domain based on speech disfluencies (Orena & White, 2015; Yoon, Jin, Brown-Schmidt, & Fisher, 2021) and so it seems possible that they might be able to extend these inferences to make judgements about others' knowledgeability.

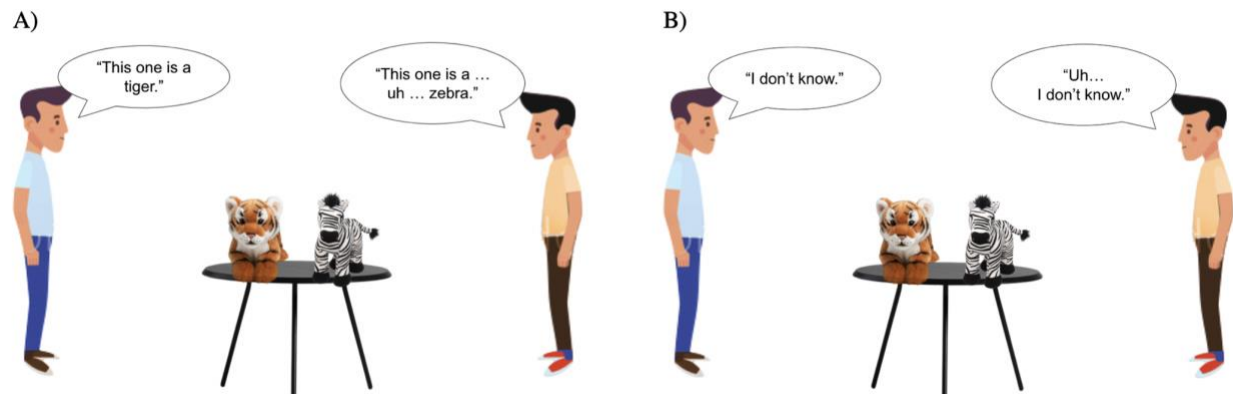


Figure 1. Example stimuli used in Experiments 1a and 1b for the Labelling condition (A) and the Ignorance condition (B).

Methods

Participants

We pre-registered a planned sample of 120 children, 60 children in each of the two conditions (Labelling vs. Ignorance). For each condition, we planned to collect data from 20 children in each of 3 pre-determined age-groups: 4-5 years-old, 6-7 years-old, and 8-9 years-old. Data were collected at a local science museum. Due to timing constraints for collecting data in a museum collection, sample demographics beyond participant gender were not collected. After the suspension of human subjects research following the COVID-19 pandemic, we were unable to complete data collection as intended with the 8-9 year-old sample. Our museum sample

included 40 4-5 year-olds (mean age = 5.03, 16 girls), 40 6-7 year-olds (mean age = 6.86, 16 girls), and 21 8-9 year-olds (mean age = 8.88, 13 girls). An additional 19 children in the 8-9 year-old age group (mean age = 8.86, 10 girls) completed a minimally adapted version of the task over Zoom with a live experimenter. More details on the general online protocol can be found in Experiment 2 (which was conducted completely online).

Procedure

Children were randomly assigned to the Labeling or Ignorance condition. In both conditions, children were presented with an animated story on an iPad about two speakers and two familiar stuffed animals (see Figure 1). In the story, each speaker is asked “What is this animal called?” about the animal closest to them, and the experimenter reads their replies. In the Labelling condition, one speaker fluently labels their animal (“This one is a tiger”) and the other speaker disfluently labels the other animal (“This one is a... uh... zebra”). In the Ignorance condition, both speakers produce non-answers, but one does so disfluently (saying, “Uh... I don’t know”). Children were then asked a domain-wide knowledge question: “Who do you think knows more about animals– this person or this person?” while the experimenter pointed to each speaker. If children failed to choose one of the speakers (e.g., saying “both”) or failed to respond within 5 seconds, the experimenter repeated the question one time.

Children completed two trials with different animals and different speakers. For the animals, we selected a tiger and a zebra for the first trial, and a cow and a pig for the second trial. For these animals, we deliberately selected familiar animal labels with no readily available alternative basic labels (e.g., dog, doggie, puppy). Across participants, we counterbalanced the speaker order (whether the first speaker was fluent or disfluent) and the location of the two animals, yielding 4 counterbalanced orders per condition. Note that this also counterbalances

which animal the disfluency is paired with across participants and which animal is referenced first. In the second trial, speaker order was always the reverse of the first trial.

Results

For Experiment 1a, our central predictions were that children would select the fluent speaker as more knowledgeable at above chance rates in the Labelling condition, and that this rate would be significantly higher than in the Ignorance condition. To compare choices across conditions, we used a mixed effects logistic regression predicting speaker choice by condition and age (continuous), with random effects of subject. There was a significant effect of condition, such that children were significantly more likely to choose the fluent speaker as more knowledgeable in the Labelling condition, compared with the Ignorance condition ($\beta = -1.79, p < .001$). There was also a significant main effect of age ($\beta = 0.22, p = 0.02$).

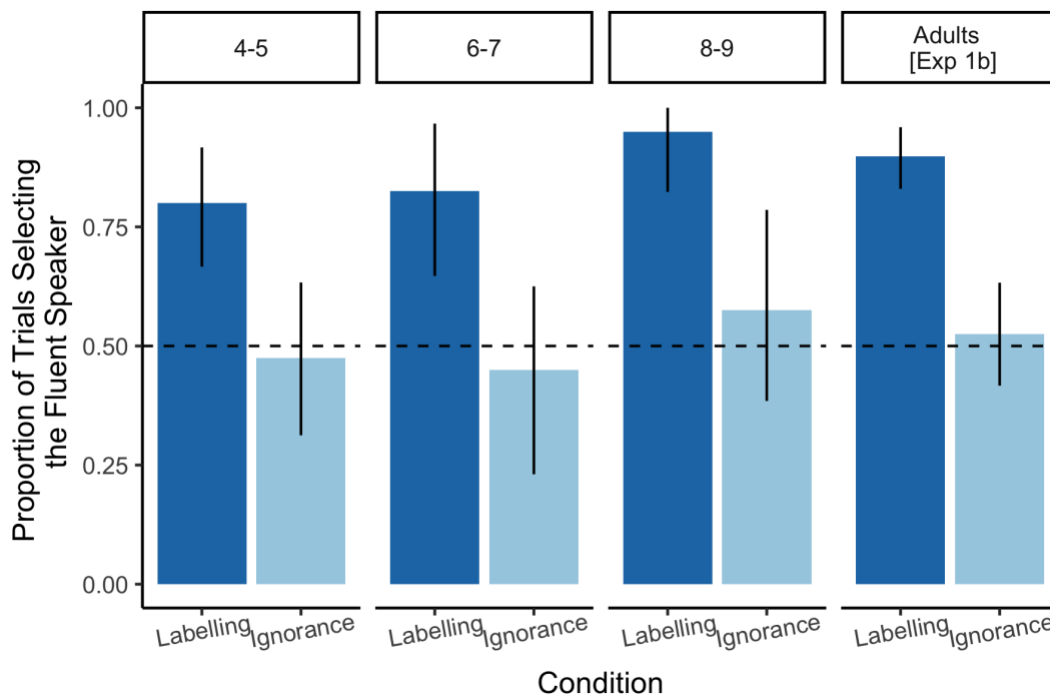


Figure 2. Age-binned results from Experiments 1a and 1b, with bootstrapped 95% confidence intervals (black lines). The dashed line indicates chance responding.

Even the youngest children in our sample show this pattern (see Figure 2). Looking only at the 4- to 5 year-olds in the Labelling condition, children selected the fluent speaker as more knowledgeable (mean proportion of trials = 0.80) significantly more often than chance ($t(19) = 4.49, p < .001$). Similarly, 6-7 year old children reliably selected the fluent speaker as more knowledgeable in the Labelling condition ($t(19) = 3.90, p < .001$) as did 8-9 year-old children ($t(19) = 9, p < .001$). In the Ignorance condition, 4-5 year old children's responding did not differ from chance (mean proportion of trials = 0.48), suggesting they were not reliably selecting either speaker ($t(19) = -0.29, p = .77$). Similarly, 6-7 year old children did not reliably choose either speaker in the Ignorance condition ($t(19) = -0.49, p = 0.63$), nor did 8-9 year-old children ($t(19) = 0.78, p = 0.45$).

Discussion

In Experiment 1a, children made robust, selective knowledge inferences, judging an accurate but disfluent speaker to be less knowledgeable than an accurate and fluent speaker. We see evidence that even 4-5 year old children are consistently making this inference. These results demonstrate that children are tracking speech disfluencies and using them to make social inferences about another person's knowledge from a young age, as adults do (Brennan & Williams, 1995). Children seem to understand that disfluencies can license inferences about the speaker's processes—namely that the speaker had difficulty (or at least delay) retrieving the appropriate label for the item and this implicates their likely knowledge.

Importantly, and as predicted, children were much less likely to infer that the fluent speaker was knowledgeable when they expressed ignorance; children at all ages showed no reliable speaker preference in the Ignorance condition. While children did not show a directional preference in the Ignorance condition, this pattern of results helps rule out the idea that children

are heuristically tracking disfluencies and always ascribing incompetence. If children held a heuristic that disfluent speakers are less competent broadly, then they should have said so here as well. This finding also suggests that our results cannot be explained by low-level auditory features. There is evidence that preschool age children are sensitive to auditory fluency cues, judging speakers whose utterances were inflected with white noise to be less competent than matched speakers who spoke without such background noise (Bernard, Proust, & Clément, 2014). However, such an effect cannot explain these results as the Ignorance condition contains an identical amount of disfluency, yet children make no such knowledge inferences.

There was one result that was somewhat unexpected in light of past work. Previous research suggests that adults differentiate fluent and disfluent non-answers, judging disfluent non-answers to indicate *greater* knowledgeability than fluent non-answers (Brennan & Williams, 1995). However, in the data for the ignorance condition we saw no evidence that even our oldest kids made such a differentiation. In our next experiment, we collected adult judgements in this task.

Experiment 1b

To determine the adult-like pattern of responses in this task, we collected knowledge judgments from a sample of adults online. While past studies have addressed similar questions in adult samples (Brennan & Williams, 1995), our task differs along a number of dimensions and so a separate baseline of adult responses is required. Additionally, while Brennan and Williams (1995) reported that adult listeners inferred that disfluent non-answers indicated more knowledge, we saw no evidence that children were making this inference in Experiment 1a. Experiment 1b provides adult comparison data for our key developmental effects from Experiment 1a, and also allows us to ask whether adults systematically infer that the disfluent

speaker is more knowledgeable in the Ignorance condition, as the results from Brennan and Williams (1995) would suggest.

Methods

Participants

120 participants were recruited from Amazon Mechanical Turk, with 60 participants in each of the two conditions (Labelling and Ignorance). Participants were paid a small reward in exchange for completing the study. Two participants had incomplete data and were excluded from the final sample.

Procedure

Experiment 1b was an online adaptation of Experiment 1a. In Experiment 1a, the experimenter read the story and speaker's utterances aloud to children. In Experiment 1b, adult participants were instead asked to read through survey slide-style Qualtrics, and then select which speaker would know more about animals. Otherwise, the stimuli and trial structure were exactly the same. As in Experiment 1a, participants completed two trials.

Results

Adult responses closely mirrored children's responses (see Figure 2). To confirm a condition-wise difference, we ran a mixed effects logistic regression predicting speaker choice by condition, with random effects of subject and trial number. There was a significant effect of condition, such that adults were significantly more likely to choose the fluent speaker as more knowledgeable in the Labelling condition, compared with the Ignorance condition ($\beta = -0.36$, $p < .001$).

Adults choose the fluent speaker as more knowledgeable in the Labelling condition (mean proportion of trials = 0.90) significantly more often than chance ($t(58) = 11.81$, $p <$

.001). In the Ignorance condition (mean proportion of trials = 0.53), adults did not reliably select either speaker ($t(58) = 0.45, p = 0.65$).

Discussion

Overall, adults (like children) judged an accurate but disfluent speaker to be less knowledgeable than an accurate and fluent speaker, but made no such inference when the speakers are claiming ignorance, which is consistent with prior work on adults' inferences in similar tasks (e.g., Brennan & Williams, 1995). However, prior work had also found that adults actually associate disfluency with *greater* knowledge for non-answers (Brennan & Williams, 1995). It has been argued that disfluency before a non-answer (e.g., "Um... I can't remember") may indicate an attempt to retrieve the relevant information (and thus relatively more knowledge), while a speedy, fluent non-answer may reflect certainty that the speaker does not possess the relevant information (Brennan & Williams, 1995). In our data, adults judgements do not reflect such inferences; however, our task differs along a number of key dimensions that make direct comparisons difficult (e.g., the simplicity of the information at issue, the use of audio) and that inference is not the focus of this paper. For our purposes, adults responses highlight the contextual role of fluency in judgements of knowledgeability and display strong developmental consistency across a wide age range.

Experiment 2a

Our account holds that disfluencies can be interpreted by listeners as domain general signs of processing time or conflict. As such, disfluencies should generate inferences in a variety of domains, beyond knowledgeability. In Experiments 2a and 2b, we ask how adults and children use speech disfluencies to infer an agent's preferences.

To examine children's ability to infer preferences from disfluency, in Experiment 2b, we changed the question under discussion so that each character is asked which of the two animals they like the best. Disfluency in responding here may also license inferences about the speaker's mental state, likely not because of difficulty retrieving the appropriate label (as in the Experiments 1a and 1b), but instead because the speaker is experiencing conflict between the two options, delaying their response time. Thus, a speedy, fluent response may indicate a strong and decisive preference (similar to the degree of agreement inference made by adults in Roberts et al. (2011), see also Gates, Callaway, Ho, and Griffiths (2021) for similar non-linguistic findings). When two speakers both state the same preference, but one does so disfluently, we predicted that, children would infer that the disfluent speaker had a relatively weaker preference. This would demonstrate that children can use the same disfluency cue to make inferences about one's mental state in a domain beyond knowledge.

We further expected that this situation might license an additional inference, generalizing to an unmentioned (and thus dispreferred) co-present item. Perhaps the disfluent speaker would be seen as having a weaker preference for the preferred animal (as described above) because they have relatively split preferences— maybe the speaker was conflicted and even considered picking the alternative. To test for such reasoning, we asked participants to also report who likes the unmentioned animal more (i.e. the dispreferred item). If participants reason in this way, they should infer the disfluent speaker may have a relatively stronger preference for this item. We thus predicted that adults (and possibly children) would make opposite inferences about a disfluent speaker's relative preference when asked about the item that the speaker selected and the item they did not select. Note, we hypothesized that inference about the unmentioned animal might be difficult for our youngest children, based on some evidence that children as young as

infancy seem better able to represent preference than dispreference (Feiman, Carey, & Cushman, 2015).

Importantly, our logic outlined above should not apply to the knowledge inferences made in labelling contexts explored in Experiments 1a and 1b. In the case of labelling, if anything, the speaker who disfluently labels a given animal should be seen as knowing less in the domain as a whole, including unmentioned animals. Indeed, our data from Experiments 1a and 1b ask children and adults to make domain-wide knowledge inferences— participants were asked “Who knows more about animals?”. At the very least, there is no reason to expect the disfluent speaker would somehow be *more* knowledgeable about the unmentioned item (whereas for preferences we do predict the disfluent speaker will show a stronger relative preference for the unmentioned item). In this way, across the two conversational contexts, we expect that the meaning of the disfluency should prompt quite distinct inferences about how the speakers view an unmentioned item.

Such a pattern of selective generalization to unmentioned items would be strong evidence that children are interpreting disfluencies in a richly contextual manner. To first establish this pattern of inferences, we collected judgments in this task from a sample of adults. In Experiment 2a, we sought to confirm that adults would use disfluency to make the hypothesized selective generalization inferences. Additionally, to our knowledge past studies with adults have not investigated the dispreference inference we aimed to capture in our task. Thus in Experiment 2a, we ask whether (1) adults use disfluency to reason about a speaker’s preference as well as dispreference, and (2) whether they generalize the meaning of a disfluency differently across our Preference and Labelling contexts.

Methods

Participants

60 participants were recruited from Amazon Mechanical Turk. Participants were paid a small reward in exchange for completing the study.

Procedure

The general procedure for Experiment 2a was largely matched to the previous experiments. The central change of interest was to shift to the domain of preferences, rather than knowledge. As with Experiment 1b, adult participants were asked to read through survey slide-style Qualtrics, and then answer questions.

Preference Trial

In Experiment 2a, two characters were introduced one at a time, and each character was given an identifiable color by which they were referred, for example “the blue person” (see Figure 3). In Experiment 2a, each speaker entered the scene independently, and was alone when asked a question. Viewing the tiger and the zebra, each speaker was asked, “Which of these animals is your favorite? Which one do you like the best?” Both speakers independently stated a preference for the same animal, but one did so disfluently (e.g., saying “Um... the [tiger] is my favorite”). We had speakers respond one-at-a-time and alone to reduce inferences of possible social motivations for stating the same preference (or stating that preference slowly). At the end of the trial both speakers were then brought back and participants were reminded what each speaker said. This ensured that both speakers were onscreen when the target questions were asked. Across participants, we counterbalanced whether the first or second speaker was the disfluent speaker, and which animal was preferred.

Participants were then asked three questions: a preference (mentioned animal) question: “Who do you think likes the [tiger] more– the blue person or the green person?”, a distractor

question: “Who do you think is better at playing basketball– the blue person or the green person?”), and a preference (unmentioned, dispreferred animal) question: “Who do you think likes the [zebra] more– the blue person or the green person?” The order of the two target questions was counterbalanced across participants. The distractor question was always asked second to ensure participants were not merely switching responses across questions (this was primarily done for our child participants in Experiment 2b, but was kept the same for consistency in our adult sample). Participants completed one Preference trial.



Figure 3. Example stimuli used in Experiment 2 for the Preference trial. In this example, the tiger would be the ‘mentioned’ animal, and the zebra would be the ‘unmentioned’ animal. Note that in Experiment 2 speakers entered one-at-a-time and then shown together at test.

As an initial comparison, we included an additional Labelling trial always after the key Preference trial described above. The structure of the Labelling trial was adapted slightly from Experiments 1a and 1b to better mirror the Preference trial described above. As in Experiments 1a and 1b, the characters in the story were asked, “What is this animal called?” but both were asked about the same animal and respond while alone (as in the Preference Trial). Each speaker

labelled the animal accurately, but one did so disfluently (e.g., saying “This one is a... um... pig.”). Then, both speakers were shown and participants were reminded who said what.

Participants were then asked three questions: a mentioned knowledge question: “Who do you think knows more about [pigs]?”, a distractor question: “Who do you think is better at playing soccer?”, and an unmentioned knowledge question: “Who do you think knows more about [cows]?” The order of the mentioned and unmentioned knowledge questions was counterbalanced across participants. The distractor question was always asked second to ensure participants were not merely switching responses across questions. Note that unlike Experiments 1a and 1b, these knowledge questions ask participants to evaluate knowledge at the item-level, rather than the domain level. This change allowed questions to be more analogous to the Preference trial, and allowed us to look at potential differences in how these inferences extend to the unmentioned items specifically. Participants completed one Labeling trial, always after the key Preference trial.

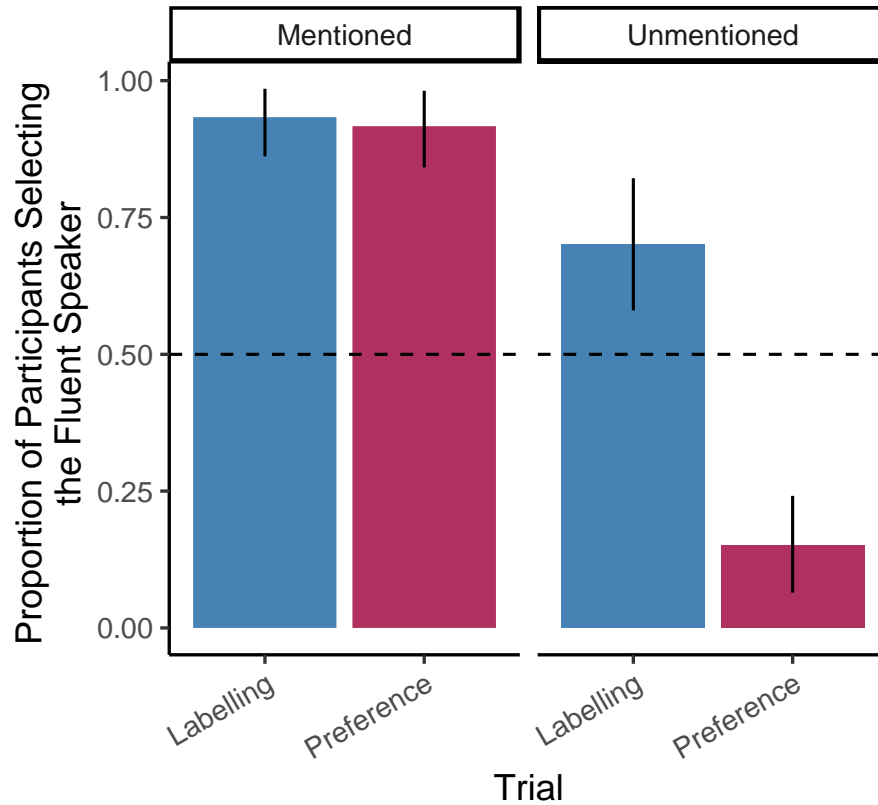


Figure 4. Adults' responses from Experiment 2a with bootstrapped 95% confidence intervals (black lines). The dashed line indicates chance responding. These data illustrate the pattern of selective generalization to the unmentioned item.

Results

Preference

We first analyze adults' judgements from the Preference trial (see Figure 4). When asked which speaker likes the preferred animal more, adults were significantly more likely to select the fluent speaker (mean proportion of adults = 0.92) than the disfluent speaker ($t(59) = 11.58, p < .001$). But, as predicted, when asked which speaker likes the dispreferred animal more, adults were significantly less likely to select the fluent speaker (mean proportion of adults = 0.15) than the disfluent speaker ($t(59) = -7.53, p < .001$).

Labelling

Next, we analyze adults' judgements from the Labelling trial. When asked which speaker knows more about the mentioned animal species, adults were significantly more likely to select the fluent speaker (mean proportion of adults = 0.93) than the disfluent speaker ($t(59) = 13.34$, $p = < .001$). This finding mirrors our results from Experiment 1b, but here using an item-specific measure of knowledge, and is consistent with the prior literature (Brennan & Williams, 1995). We further expected that adults would extend this reasoning beyond the item under discussion, judging that a fluent speaker may also know more about the unmentioned co-present item. Indeed, when asked who knows more about the *unmentioned* animal species, adults were significantly more likely to select the fluent speaker as more knowledgeable (mean proportion of adults = 0.70; $t(59) = 3.35$, $p = .001$).

Discussion

In the key Preference trial, adults inferred that a disfluent speaker likely has a weaker preference for the item under discussion, but also inferred that speaker would have a relatively stronger preference for an unmentioned alternative (also see Gates et al., 2021). This result was aligned with our hypothesis rooted in the idea that people are making inferences about a choice conflict for the disfluent speaker. If the participant inferred that the disfluent speaker was conflicted about which animal was their favorite, then they should think both that the disfluent speaker has a weaker preference for the selected animal than the fluent speaker, and vice versa for the unmentioned animal. In contrast, for the Labelling trial, adults inferred that a disfluent speaker would likely have less knowledge about the animal under discussion, and also extended that reasoning to think they might also have less knowledge about a related, unmentioned animal type. Overall, these results provide strong support for the selective generalization inference that

we predicted, with adults drawing opposite inferences about the unmentioned animals across the two conversational contexts. To our knowledge, this research question has not been previously explored in adults and these results help further demonstrate the flexibility and contextuality of mature, adult reasoning about the implications of disfluencies. In Experiment 2b, we examine the development of children's responses on a closely matched task to ask how they use disfluency to reason about speaker preference, and whether we see this hallmark of selective generalization across the two contexts.

Experiment 2b

Having found support for our predictions about how adults make inferences about disfluencies in relation to preferences and labeling, we next explored whether 4- to 9-year-old children make these inferences and how these inferences develop over this age range. We hypothesized that the preference inferences about the unmentioned item might be difficult for our youngest children, because it requires reasoning about dispreference and there is some evidence that children as young as infancy seem better able to represent preference than dispreference (Feiman, Carey, & Cushman, 2015).

Methods

Participants.

We recruited a pre-registered sample of 60 children to run in Experiment 2, with 20 children in each of 3 pre-determined age-groups: 4-5 years-old, 6-7 years-old, and 8-9 years-old. Due to overrecruitment, our final sample included 64 children: 21 4-to-5-year-olds (mean age = 5.00, 9 girls), 21 6-to-7-year-olds (mean age = 6.90, 10 girls), and 22 8-to-9-year-olds (mean age = 8.85, 11 girls). These data were collected online via Zoom with a live experimenter.

Participating families were largely recruited via a participant database of Chicagoland families who have previously participated in in-person research studies.

Procedure

Experiment 2b was largely matched to Experiment 2a. While adult participants were asked to read through survey slide-style Qualtrics in Experiment 2a, in Experiment 2b, the experimenter read the story and speaker's utterances aloud to children. Otherwise, the stimuli and trial structure were exactly the same as described in Experiment 2a.

We also made a few alterations to the task structure from Experiment 1a to facilitate children's responses. In Experiment 1a, pointing was the dominant response, and that behavior can be difficult to reliably capture in remote testing. In Experiment 2b, we used identifiable colors to refer to each of the speakers throughout the story and questions. This change was made to facilitate children's ability to respond verbally in Experiment 2b when asked to select a character, rather than pointing.

Results

We first analyze children's judgements from the Preference trial (see Figure 5). To compare choices across question types (mentioned vs. unmentioned), we used a logistic regression predicting speaker choice by condition and age (continuous). There was a significant effect of question, such that children were significantly more likely to choose the fluent speaker as preferring the mentioned item, compared with the unmentioned item ($\beta = -7.68, p < .001$). There was also a significant main effect of age ($\beta = -0.70, p = < .01$), and a significant interaction ($\beta = 1.50, p = < .001$). We next examine these developmental shifts in children's responses.

We analyze responses separately for each of our pre-determined age groups. For 4-5 year old children, when asked which speaker likes the preferred animal more, children did not reliably select the fluent speaker (mean proportion of children selecting fluent = 0.43) or the disfluent speaker ($t(20) = -0.65, p = .526$). In contrast, 6-7 year old children consistently selected the fluent speaker as having a stronger preference (mean proportion of children selecting fluent = 0.85) than the disfluent speaker ($t(19) = 4.27, p < .001$). 8-9 year old children showed this same pattern of selecting the fluent speaker as having a stronger preference for the mentioned item (mean proportion of children selecting fluent = 0.95) than the disfluent speaker ($t(21) = 10, p < .001$).

We see a similar age-related pattern for children's judgments about the dispreferred, unmentioned animal. For 4-5 year old children, when asked which speaker likes the unmentioned animal more, children did not reliably select the fluent speaker (mean proportion of children selecting fluent = 0.55) or the disfluent speaker ($t(19) = 0.44, p = 0.66$). In contrast, 6-7 year old children were more likely to select the disfluent speaker as having a stronger preference for the unmentioned item (mean proportion of children selecting disfluent = 0.76) than the fluent speaker ($t(20) = -2.75, p = .01$). 8-9 year old children showed this same pattern of selecting the disfluent speaker as having a stronger preference for this unmentioned item (mean proportion of children selecting disfluent = 0.95) than the fluent speaker ($t(21) = -10, p < .001$).

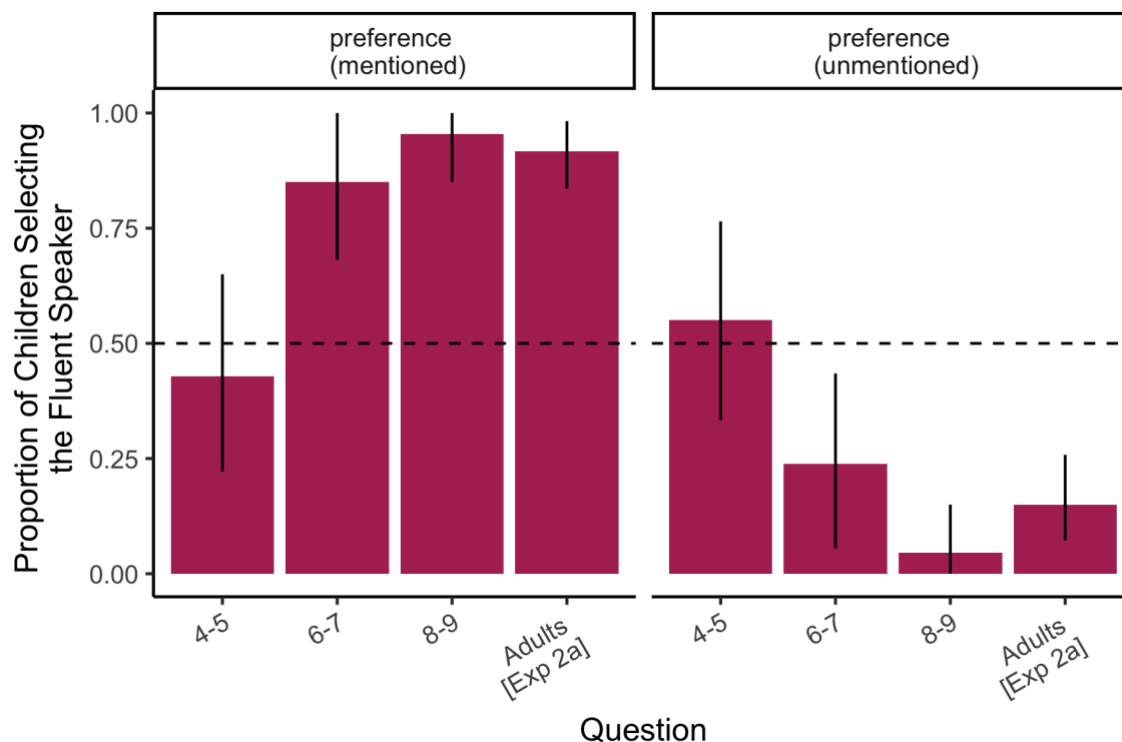


Figure 5. Preference trial results from Experiments 2a and 2b with bootstrapped 95% confidence intervals (black lines). The dashed line indicates chance responding.

Next, we analyze children’s judgements from the Labelling trial. Our key question was whether children’s responses differed reliably for the unmentioned items across the Labelling and Preference trials. To test this difference, we used a logistic regression predicting speaker choice by condition and age (continuous). Contrary to our expectations, there was no significant effect of condition on children’s response for the unmentioned items ($\beta = -2.53, p = 0.18$). There was a significant effect of age ($\beta = -0.70, p < 0.01$). There was no significant interaction between age and condition ($\beta = 0.38, p = -0.19$). We next analyze children’s responses separately for each of our pre-determined age groups to further investigate the effect of age, and the unexpected null effect of condition on children’s responses about the unmentioned item.

First looking at children's response to the mentioned items, we replicate our results from Experiment 1a. When asked which speaker knows more about the mentioned animal, even children as young as 4-5 years old reliably selected the fluent speaker (mean proportion of children selecting fluent = 0.76) more than the disfluent speaker ($t(20) = 2.75, p = 0.01$). This pattern remained consistent for 6-7 year old children ($t(20) = 9.50, p < .001$) and 8-9 year old children ($t(21) = 6.52, p < .001$).

Next we turn to children's responses about the unmentioned item in the Labelling trial. We predicted that when asked who knows more about the unmentioned animal, children (at least older children) would be at chance or even favor the fluent speaker. However, contrary to our expectations and adult data, children overall were actually significantly more likely to select the *disfluent* speaker as more knowledgeable about the unmentioned animal (mean proportion of children selecting disfluent = 0.75; $t(63) = -4.58, p < .001$). When asked which speaker knows more about the unmentioned animal, 4-5 year old children did not reliably select the fluent speaker (mean proportion of children selecting fluent = 0.38) or the disfluent speaker ($t(20) = -1.10, p = 0.29$). However, 6-7 year old children were more likely to select the disfluent speaker as more knowledgeable about the unmentioned animal (mean proportion of children selecting disfluent = 0.76) than the fluent speaker ($t(20) = -2.75, p = 0.01$). 8-9 year old children also showed this same pattern of being more likely to select the disfluent speaker as more knowledgeable about the unmentioned animal (mean proportion of children selecting disfluent = 0.86) than the fluent speaker ($t(21) = -4.86, p < .001$).

Discussion

Replicating and extending our results from Experiment 1a, we find that children use disfluencies to infer both others' knowledge and preferences. We again found that children at all

ages tested inferred that a fluent speaker was more knowledgeable about an item that they labelled, conceptually replicating our in-person results from Experiment 1a in paradigm adapted for remote testing. By 6- to 7-years-old, but not younger, children reliably inferred that someone who is able to fluently state their preferred item likely has a stronger preference for that item. The fact that children made these inferences about both others' knowledge and preference is consistent with an account where children reason about as signals to mental processing, with contextualized interpretation.

Our data also suggest the 4–5 year-old children may not be drawing reliable preference inferences based on speaker fluency. This marks a potentially interesting contrast to our findings from Experiment 1a (also replicated here in the Labelling trial) where 4-to-5-year-olds are able to infer knowledgeability in a closely matched task. One possibility is that connection between disfluency and knowledge emerges before it is connected to other domains like preference, at least in this task.

The data on children's inferences about the unmentioned animals are more difficult to interpret here. Children by 6-7 do reliably select the disfluent speaker as having a stronger preference for the unmentioned object, as we predicted and as adults did. However, we see a similar, unpredicted pattern for children's knowledge inferences about the unmentioned animal. While children made domain-wide competence inferences in Experiment 1b (reliably reporting that the fluent speaker knows more about the animals in general), they seem to make the opposite inference when asked about another, unmentioned animal here. That is, they are judging that someone who is disfluent at labelling one animal is *more* knowledgeable about another, unmentioned animal. This response is counter to our hypotheses, as well as our adult findings from Experiment 2a. Together, these data make the dispreference measure difficult to interpret,

and are inconsistent with children doing the selective generalization we observed in adults in Experiment 2a.

However, there are some aspects of our design that make it difficult to interpret children's responding on the knowledge trials. In this experiment, the Preference trial was our focus, thus the Knowledge trial always came after the Preference trial. It is possible that the pattern of responses evident in Preference trial carried over to the subsequent Labelling trial and affected children's knowledge judgements as well. That is, after stating that a person had a stronger preference for an unmentioned animal, they may have felt compelled to be consistent and say that the speaker also has more knowledge about an unmentioned animal. Experiment 3 tests this question in a fully between-subjects design to rule out any possible carry-over effects, and better test for the development of the selective generalization inferences.

Experiment 3

Experiment 2b suggests that children are able to extend their reasoning about speech disfluencies to the domains of both knowledge and preference, at least by age 6-to-7. As a secondary goal, Experiment 2b was also meant to test for selective generalization of these inferences across our two conversational contexts (Preference and Labelling). As we had predicted, adults showed a pattern of selective generalization in their inferences, using disfluencies to infer a stronger preference for an unmentioned animal, but not more knowledge of an unmentioned animal (see Figure 4). However, children showed no such evidence of selective inferences, with older children instead inferring both a stronger preference for and more knowledge about an unmentioned animal.

Experiment 3 was designed to replicate our major results from Experiment 2b and to examine whether possible carryover effects might have influenced children's responses in the

Labelling trial of Experiment 2b due to the fixed trial order. Thus, to better test selectivity in children's reasoning across these two contexts, Experiment 3 utilizes a fully between-subjects design. The key question is whether children generalize the meaning of a disfluency differently (to unmentioned alternatives) depending on the conversational context.

Methods

Participants

We recruited a pre-registered sample of 120 children to run in Experiment 3, with 20 children in each of 3 pre-determined age-groups: 4-5 years-old, 6-7 years-old, and 8-9 years-old. As in Experiment 2, these data were collected online via Zoom with a live experimenter. 6 participants were excluded and resampled due to experimenter error ($n = 2$), technical difficulties ($n = 3$), and interference ($n = 1$). Our final sample ($n = 121$) included 40 4-to-5-year-olds (mean age = 4.96, 24 girls), 41 6-to-7-year-olds (mean age = 6.99, 16 girls), and 40 8-9-year-olds (mean age = 8.92, 22 girls). Participating families were largely recruited via a participant database of Chicagoland families who have previously participated in in-person research studies.

Procedure

Experiment 3 was largely identical to Experiment 2b, but utilized a between-subjects design such that children were randomly assigned to either the Labelling or Preference condition. In each condition, participants completed two trials, with different speakers and animals across trials.

As in Experiment 2b, participants were asked three questions during each trial. Children were asked two target questions (one about the mentioned animal and one about the unmentioned animal), and an irrelevant distractor question (about ability in sports). In the Labelling condition,

children were asked “Who do you think knows more about [tigers]?” for both the mentioned and unmentioned animals. In the Preference condition, children were asked “Who do you think likes [tigers] more?” for both the mentioned and unmentioned animals. The order of the two target questions was counterbalanced across participants, while a distractor question was always asked second– to minimize the use of a side-switching strategy across trials.

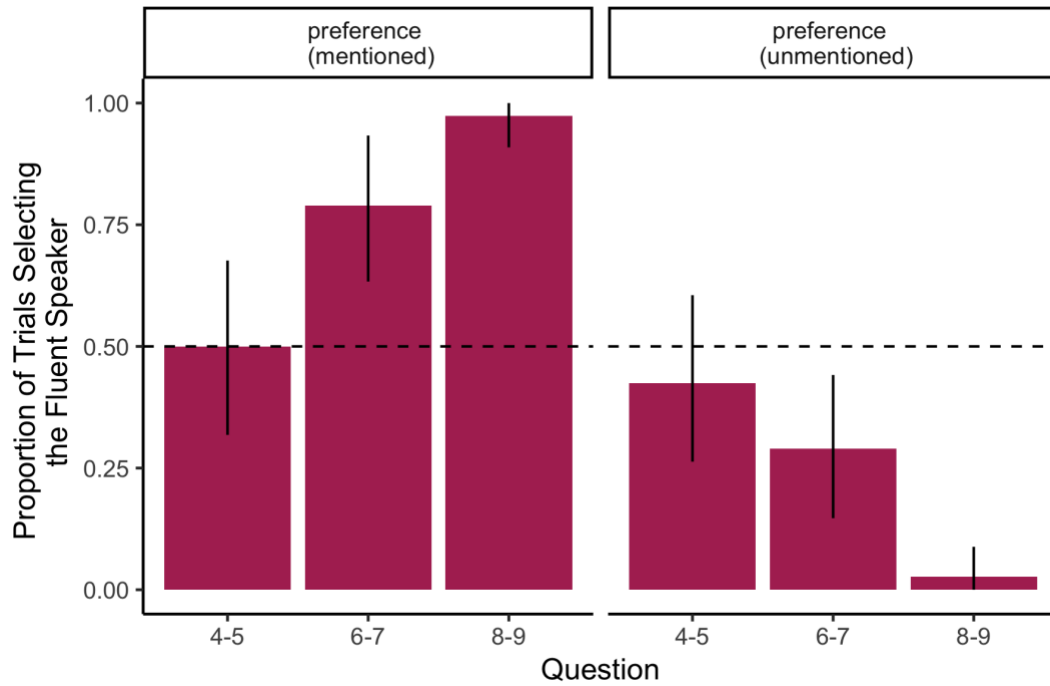


Figure 6. Preference condition results from Experiment 3 with bootstrapped 95% confidence intervals (black lines). The dashed line indicates chance responding.

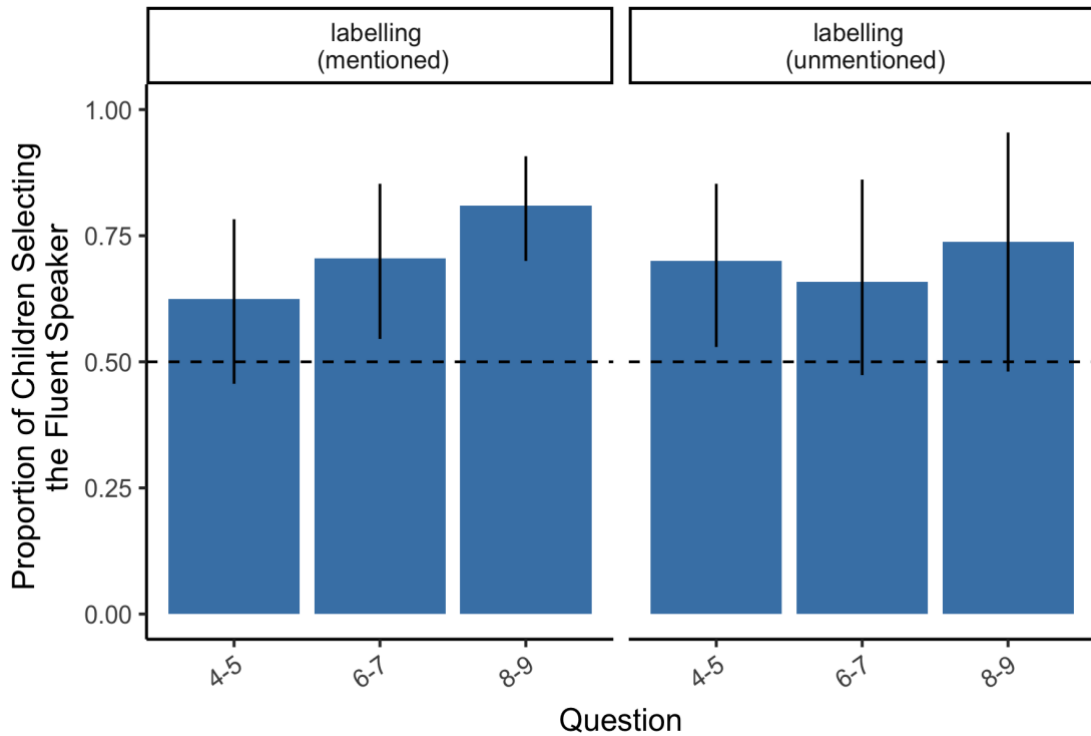


Figure 7. Labelling condition results from Experiment 3 with bootstrapped 95% confidence intervals (black lines). The dashed line indicates chance responding.

Results

We first analyze children’s judgements from the Preference condition (see Figure 6). To compare choices across question types (mentioned vs. unmentioned), we used a logistic regression predicting speaker choice by condition and age (continuous). There was a significant effect of question, such that children were significantly more likely to choose the fluent speaker as preferring the mentioned item, compared with the unmentioned item ($\beta = -7.21, p < .001$). There was a significant effect of age ($\beta = -0.69, p < .001$), and a significant interaction between age and question ($\beta = 1.45, p < .001$). We next examine these developmental shifts in children’s responses.

For 4-5 year old children, when asked which speaker likes the preferred animal more, children did not reliably select the fluent speaker (mean proportion of trials selecting fluent = 0.50) or the disfluent speaker ($t(19) = 0, p > .999$). In contrast, 6-7 year old children consistently selected the fluent speaker as having a stronger preference (mean proportion of trials selecting fluent = 0.79) than the disfluent speaker ($t(18) = 3.64, p = 0.002$). 8-9 year old children showed this same pattern of selecting the fluent speaker as having a stronger preference (mean proportion of trials selecting fluent = 0.97) than the disfluent speaker ($t(18) = 18.00, p < .001$).

We see a similar age-related pattern for children's judgments about the dispreferred, unmentioned animal. For 4-5 year old children, when asked which speaker likes the unmentioned animal more, children did not reliably select the fluent speaker (mean proportion of trials selecting fluent = 0.42) or the disfluent speaker ($t(19) = -0.83, p = 0.42$). In contrast, 6-7 year old children consistently selected the disfluent speaker as having a stronger preference for the unmentioned animal (mean proportion of trials selecting disfluent = 0.71) than the fluent speaker ($t(18) = -2.65, p = .016$). 8-9 year old children showed this same pattern of selecting the disfluent speaker as having a stronger preference for the unmentioned animal (mean proportion of trials selecting disfluent = 0.03) than the fluent speaker ($t(18) = -18.00, p < 0.001$).

We next turn to children's judgements from the Labelling condition (see Figure 7). Our key question was whether children responses differed reliably for the unmentioned items across the Labelling and Preference conditions -- consistent with our hypothesized pattern of selective generalization. To test this difference, we used a logistic regression predicting speaker choice by condition and age (continuous). There is a significant effect of condition on children's responses about the unmentioned items ($\beta = -2.92, p < .05$). There was also a significant effect of age

($\beta = 0.69, p < .001$), and a significant interaction between age and question ($\beta = 0.676, p = .001$). We next analyze children's responses separately for each of our pre-determined age groups to further investigate the effect of age, and the effect of condition on children's responses to the unmentioned item.

For 4-5 year old children, when asked which speaker knows more about the mentioned animal, children did not reliably select the fluent speaker (mean proportion of trials selecting fluent = 0.62) or the disfluent speaker ($t(19) = 1.56, p = .135$). 6-7 year old children consistently selected the fluent speaker as more knowledgeable about the mentioned animal (mean proportion of trials selecting fluent = 0.70) than the disfluent speaker ($t(21) = 2.61, p = .016$). 8-9 year old children showed this same pattern of selecting the fluent speaker as having more knowledgeable (mean proportion of trials selecting fluent = 0.81) than the disfluent speaker ($t(20) = 5.70, p < .001$).

We predicted that when asked who knows more about the unmentioned animal, children would be at chance or even favor the fluent speaker, as adults do in Experiment 2a. In Experiment 3, children overall were significantly more likely to select the fluent speaker as more knowledgeable even about the unmentioned animal (mean proportion of trials selecting fluent = 0.70), just like adults ($t(62) = 3.34, p = .001$). This contrasts with our unexpected findings in children's responses in Experiment 2b, and we return to these findings in our discussion. For 4-5 year old children, when asked which speaker knows more about the unmentioned animal more, children were more likely to select the fluent speaker (mean proportion of trials selecting fluent = 0.70) than the disfluent speaker ($t(19) = 2.37, p = 0.03$). 6-7 year-old children showed no robust pattern for selecting the fluent speaker or the disfluent as more knowledgeable about the unmentioned animal (mean proportion of trials selecting fluent = 0.66, $t(21) = 1.58, p = 0.129$).

8-9 year-old children were marginally more likely to select the fluent speaker as more knowledgeable about the unmentioned animal (mean proportion of trials selecting fluent = 0.74) than the disfluent speaker ($t(20) = 1.94, p = 0.066$).

Discussion

Consistent with our hypotheses, we replicated our previous results and also found that children made different inferences about the meaning of disfluency when asked about unmentioned items, demonstrating the selective generalization inference we saw in adults. Children ages 6-9 inferred a disfluent speaker has a relatively stronger preference for an unmentioned item (in the Preference condition), which is consistent with our predictions and the results of Experiment 2b. That is, when someone seemed conflicted about picking their favorite animal between two options, children by age 6-7 inferred that speaker must have a stronger relative preference for the animal they did not pick, compared with a speaker who fluently selected their favorite. Importantly, we also found that they do not expect a disfluent speaker to be more *knowledgeable* about an unmentioned item (in the Labelling condition). Recall that the results for knowledge about unmentioned items observed in Experiment 2b showed older children unexpectedly selecting the disfluent speaker as more knowledgeable, which we speculated may have been caused by carryover effects based on the within participant design in Experiment 2b. Our between participant design confirmed this speculation; we see no evidence of that inference in our between participant design in Experiment 3, which provides a more controlled and better powered test of this question. Indeed, if anything, children inferred that the disfluent speaker would be less knowledgeable about a related, unmentioned animal type than the fluent speaker (similar to our results from Experiment 1a). In sum, the results of Experiment

3 provide strong evidence that children are interpreting the social meanings of disfluencies in an inferential, contextual manner.

While the results in Experiment 3 are broadly consistent with the results from our previous experiments, we do see slightly different age-specific effects in our Labelling condition compared with Experiments 1a and 2b. In our previous studies, children as young as 4-5 years-old have consistently inferred that a fluent speaker is more knowledgeable, but here show no significant knowledge inference for the mentioned item. However, it is worth noting that younger children's responses here are in the same direction as in our previous studies and the younger children do make a knowledge inference for the unmentioned item in line with responses of older children and adults. Thus, on balance, across the three studies there seems to be consistent evidence that young children can make inferences about knowledge based on speech disfluencies, while we have seen no evidence that children at this age can make inferences about preference based on speech disfluencies.

General Discussion

Across 3 experiments, children draw inferences about another agent's mental states based solely on the disfluencies in their speech, and these inferences are flexible and context-sensitive. We see consistent evidence that even young children infer that an accurate, but disfluent speaker might be less knowledgeable about the topic at hand (Experiments 1a and 2b). By age 6, children similarly use disfluencies to infer the relative preferences of a speaker (Experiments 2b and 3), and they understand the meaning of a disfluency might generalize to unmentioned items differently depending on the domain in question (Experiment 3). In sum, these findings suggest that disfluencies may serve as powerful cues to a speaker's mental processes, with broad implications.

These studies add to the rich literature on children’s ability to infer knowledge and engage in selective social learning (e.g., Koenig & Harris, 2005). In our studies, children made selective knowledge judgements based on disfluencies, even while holding speaker accuracy constant. 4-5 year-old children (as young as we tested) made these contextual knowledge inferences at rates similar to adults. Our findings reveal that even our youngest children are not merely responding heuristically to the perceived confidence of a speaker, but instead flexibly reasoning about the contextual meaning of a disfluency. While prior work demonstrates that children as young as 2 are sensitive to an agent’s confidence (Birch et al., 2010) and prefer to learn words from a fluent speaker (White et al., 2020), our results cannot be explained by a simple confidence-is-preferred heuristic. Such a heuristic would struggle to account for children’s inferences in Experiment 1a, where we see that children do not prefer a fluent speaker in the Ignorance condition. Instead, our results suggest that children may understand something about the speech production process itself, and use that knowledge to reason contextually about the meaning of speech disfluencies.

Disfluencies do not only reflect knowledge; their meaning and social implications are contextually defined and can indicate a range of mental states and processes—some of which may be more difficult for younger children. Children age 4-5 seem to infer a disfluent speaker might be less knowledgeable (Experiments 1a, 2b and for some measures in Experiment 3), but do not extend this inference to the preference domain until age 6-7 (Experiments 2b and 3). 4- to 5-year-old children therefore also show no evidence for the kind of selective generalization evidenced by older children—they do not make the inference about preference for the mentioned item, and they also do not make inferences about preference for the unmentioned item (Experiments 2b and 3). We did not predict that preferences per se would be difficult for young

children. We predicted that if anything young children might struggle with a dispreference inference—namely that a speaker who fluently states their preference might more strongly *dislike* the unmentioned (and thus dispreferred) object, compared with a disfluent speaker (Feiman, Carey, & Cushman, 2015). However, we saw no such evidence and instead both the preference and dispreference inferences seemed to develop by age 6-7—children inferred both that a fluent speaker had a stronger preference for their preferred item, whereas the disfluent speaker had a stronger preference for the dispreferred item. Instead, our results present the intriguing possibility that children may more easily connect fluency and knowledge, compared with fluency and preference. However, we note that this specific result was not predicted and subsequent experiments are needed to test the robustness of this developmental change.

Some limitations of the current studies should be noted. It should be noted that the placement of the disfluency differed across conditions. In the Labelling conditions, the disfluency was always mid-utterance, immediately preceding the target noun. This choice was made to simulate a natural lexical retrieval process. However, in the Ignorance condition (Experiment 1) and the Preference conditions (Experiments 2 and 3), the disfluency was always utterance-initial, as was the case in related adult work on agreement and honesty (e.g., Fox Tree, 2002; Roberts et al., 2011). While such placement differences are unlikely to account for the full pattern of condition differences here, future experiments should match the placement of the disfluency or manipulate it directly. Additionally, these experiments all manipulated the presence of filled pauses (“umm” and “uh”), rather than other kinds of disfluencies. While not tested here, our account is not specific to filled pauses, but instead predicts that children should also derive similar inferences for other types of disfluencies, so long as those disfluencies implicate processing difficulty. Indeed, prior adult work has shown similar effects with silent pauses (e.g.,

Fox Tree, 2002; Roberts et al., 2011). There are practical difficulties in effectively testing some disfluency types (such as silent pauses) in a remote Zoom testing environment, where such pauses may be seen as technological glitches or lagging. However, future work could gain much from comparing and contrasting various types of disfluencies, as adult work has done (e.g., Fox Tree, 2002).

Adults derive a range of social inferences based on speech delay and disfluency (e.g., Brennan & Williams, 1995; Fox Tree, 2002; Roberts et al., 2011), and these inferences are likely built on shared underlying general inferences they make about processing delays. Speech disfluencies are one type of delay that then triggers broader inferences about others' minds. The applications of this underlying inference (e.g., to knowledge, willingness, comfort, and more) likely come from a contextualized interpretation of a shared broader principle (i.e. what is most likely slowing someone down in this situation). For example, after hearing someone slowly assent to a request, people may detect the difficulty in producing the response and then search for contextual cues to explain such difficulties—e.g., politeness considerations or avoidance of a dispreferred response (Roberts et al., 2011). In other words, these delays often acquire social meaning by integrating information from the conversational context to determine why a speaker likely paused in this case (Fox Tree, 2002).

While the current experiments suggest children may be able to reason about the production process that generated an utterance, future experiments should test just how richly young children are able to model a speaker's production process. For instance, adults seem to weight their inferences based on the length of a delay—inferring someone who pauses for longer before assenting is even less willing (e.g., Roberts et al., 2011). If children similarly titrate their judgements based on delay length, this would further support an inferential account of children's

responses wherein children are reasoning about the underlying production process, rather than relying on learned heuristics in response to a cue. This present work marks a crucial first step for understanding how children use disfluency to reason about other people, and prompts a number of interesting questions for future research.

In exploring these broader principles, much could be gained from jointly considering children's abilities to reason about other people based on timing more broadly (i.e. outside of conversation). In the physical domain, young children believe an agent who successfully builds a tower faster than another agent building the same tower is better at building (Leonard, Bennett-Pierre, & Gweon, 2019). Relatedly, children use the speed a character solves puzzles in a story to infer competence, at least in some contexts (Heyman & Compton, 2006). For complex reasoning problems, children by age 7 seem to use response time to infer the likely mental process that generated the solution, such that a quick answer is likely a retrieved memory and a longer response time is likely an in-the-moment solution (Richardson & Keil, 2022). These last findings suggest that children are starting to understand a different aspect of delay than explored in the current work – namely that delays may sometimes reflect a kind of deliberativeness.

Accordingly, there may be instances where delays signal a kind of active processing that itself indicates knowledge of a given problem. Such reasoning is more likely for delays with complex reasoning problems, rather than the simple lexical retrieval cases studied here. Related research in moral reasoning with adults suggests that delays before making immoral decisions may reflect a kind of deliberativeness that leads adults to judge an actor as a relatively more moral person (Cricher, Inbar, & Pizarro, 2013)—at least they are the kind of person who balked at the idea of doing something wrong. Beyond conversation, actions and events themselves are profoundly

structured by time in ways even young children are likely very familiar with, and these studies suggest that timing may be an especially useful cue in the development of social reasoning.

Disfluencies are in-the-moment cues that can powerfully shape the meaning and social implications of what someone says. Indeed, even some digital assistants, such as Google's Duplex, now produce disfluencies in an effort to simulate more naturalistic speech and active processing. While these cues are doubtless helpful in language processing (Kidd et al., 2011) and language learning (White et al., 2020), this work extends such findings to suggest that these cues additionally underlie a range of mental inferences in early childhood. Disfluencies track production difficulties, and thus can provide a useful window into an agent's mental processes, distinct from the content of what is said. Our results suggest that young children may indeed understand something of the speech production process—that disfluencies reflect a kind of thinking out loud—and leverage such knowledge to reason about delays and their implications. As young children learn about the social world, tracking speech disfluencies and processing delays broadly provide a rich dataset for learning about the social world.

Chapter 2: Reasoning about others' expectations to learn stereotypes

In Chapter 1, we demonstrated how children can use conversational cues to reason about a speaker's knowledge and preferences. In Chapter 2, we expand this project in two ways to ask (1) how children might use these cues in *feedback*, and (2) what the consequences of these skills might be beyond social reasoning. Imagine a young boy expressing a gender counter-stereotypical preference (e.g., wanting to buy a Barbie doll) and his caregiver provides a permissive, gender egalitarian response. However, imagine that response comes slowly, with markers of surprise and production difficulty (e.g., "Oh! Um... Sure"). What message does that young boy really receive? In this chapter, we explore how children and adults reason about surprisal in these situations and how these cues provide data to infer speaker expectations and to learn about normative behavior (and even stereotypes).

Surprise is a basic emotion that occurs in the face of unexpectedness, and thus witnessing *others'* surprise can license inferences about others' expectations, a kind of vicarious surprise. Adults show sophisticated abilities to reason about others' emotional expressions (including surprise), rationally and flexibly inferring underlying mental states accordingly (e.g., Wu et al., 2018). Building off such work, we investigate how others' surprise might provide rich information about the structure of social expectations. For adults, reasoning about others' reactions in this way would provide crucial insights into a speaker's expectations, extant stereotyped beliefs, and even for learning norms in a new social environment (e.g., how casually to dress in a new workplace). For children, the consequences may be even more profound.

Conversations with caregivers and other adults provide a fundamental venue for children to learn about the social world, and consequently for the transmission of stereotypes. Even ostensibly well-meaning messages can often have unintended consequences, with subtle

linguistic cues highlighting stereotype information (e.g., Chestnut et al., 2021; Moty & Rhodes, 2021; Rhodes et al., 2012). For example, explicitly egalitarian statements like “Girls are just as good as boys at math” can still perpetuate gendered ability stereotypes by setting boys as the reference point (Chestnut et al., 2021).

Beyond isolated messages, feedback and responsiveness from others also holds rich social information for young children. Research has demonstrated that others’ non-verbal affect may foster stereotype transmission (Skinner et al., 2020), but we argue that others’ expressions of surprise may hold particularly stereotype-relevant information by communicating their expectations. We know that children ages 6-to-8 can use others’ marked facial expressions of surprise to derive social inferences, e.g. about another agent’s competence (Asaba et al., 2020). For example, if two children successfully score a basket, but only one’s success leaves the teacher visibly shocked (actually dropping her jaw), we can infer who is likely the better player. Others’ emotional expressions— even non-valenced reactions like surprisal— can thus convey substantive information about the social world (Asaba et al., 2020; Wu et al., 2021).

But of course subtle social information is not just written on our faces; it also leaks out through the *linguistic* channel— specifically surprisal interjections (e.g., “oh”) and disfluencies (e.g., filled pauses like “um”). Surprisal interjections definitionally index speaker expectations, and two key observations suggest disfluencies may also license inferences about a speaker’s expectations. First, decades of cognitive science experiments demonstrate that violations of expectations delay response times in both children and adults (Meyer et al., 1997; Schützwohl & Reisenzein, 1999). As a result, conversational responses may similarly be slowed following unexpected information or behavior. Second, adults interpret others’ disfluencies in contentious conversations (e.g., about gun control) as reflecting underlying discomfort with the topic at hand

and potential dishonesty (Fox Tree, 2002). Together, these findings suggest that these cues reliably co-occur with speaker surprisal and thus may lead adults and children to derive novel inferences about a speaker's underlying expectations.

To explore how these cues to speaker expectations could inform stereotype transmission, we focus on the domain of gender stereotypes as a case study (Experiments 1 and 3). While the general inferential process could support learning many kinds of expectations (as we explore in Experiment 2), the development of gender stereotypes provides an important and ecologically-valid test case. Gender stereotypes emerge early in development; as young as 3, children show robust gender stereotypes about toy preferences, and report that their parents would be less approving of playing with a counter-stereotypical toy (Eisenberg et al., 1982; Freeman, 2007). By age 6, children show gender biases in their beliefs about ability and this affects their own decisions about which opportunities to pursue (Bian et al., 2017). To be able to combat such stereotypes, we must better understand the transmission processes underlying stereotype transmission.

General Approach

In three experiments, we take a social learning approach to ask how children and adults can use linguistic cues of surprisal to reason and learn about what kinds of behaviors are expected, even when these cues leak information that is counter to the speaker's explicit messaging. In each experiment, an adult figure affirms a character's choice (e.g. "Sure, you can have that one") and shows no facial expressions of surprise (maintaining a consistent, positive facial expression). However between conditions, we vary the presence or absence of conversational markers that tip the adult's hand—indicating whether they *did* or *did not* expect the child to make such a choice.

In Experiment 1, we ask whether children use surprisal feedback to infer if a target boy's toy choice is in line with gender stereotypes. In Experiment 2, we explore this same inference in novel categories to probe how these cues could serve as a plausible mechanism for both adults and children to learn about the descriptive and normative expectations of the social world. In Experiment 3 with adults only, we connect these two experiments to ask how surprisal cues can lead adults to learn a novel gender stereotype.

Stimuli Creation

For each experiment, we followed the same general procedure to create test utterances that varied across conditions. We started by having native speakers record surprisal utterances that contained interjections and disfluencies (e.g., “Oh really? Um... Sure, honey. Uh... We can buy you that one”), reading them as naturally as possible. We then digitally removed the surprisal markers to create corresponding fluent utterances that were well matched (e.g., “Sure, honey. We can buy you that one.”). Thus, the only features that varied across test utterances was the presence or absence of interjections and disfluencies. Utterances may have included additional paralinguistic markers outside of the interjections or disfluencies themselves (e.g., rising intonation in other phrases), but this information was matched across our conditions.

Experiment 1

In a pre-registered experiment, children were shown videos in which a target boy is choosing between two gender stereotyped toy options (e.g., a doll or a truck), and his choice was ambiguous from the participant's perspective. Children then saw an adult figure respond approvingly, but either with cues to surprise (in the surprise condition) or fluently (in the fluent baseline condition). Children were then asked to infer which toy the target boy had selected. In this way, this experiment asks how children use feedback to reason about whether a choice was

expected (i.e. stereotypical) or unexpected (i.e. counter-stereotypical). The key prediction was that children would be more likely to infer the target boy had selected a girl-stereotyped toy in the surprise condition, as compared to the fluent baseline condition.

Method

Participants

We pre-registered a sample size of 120 children ages 4-to-9, with 20 children in each condition in each of three pre-registered age bins (4-5, 6-7, 8-9). Families were recruited online, primarily through a US University database of families who have expressed interest in doing research. Children completed this experiment over Zoom, interacting with a live experimenter who navigated a slide-style, animated Qualtrics survey. Based on a pre-registered exclusion criterion, children who failed to answer all of the questions were excluded and replaced (an additional 6 children).

Procedure

Participants were shown two short, animated stories that featured different protagonists and toys. Each story was about a young boy and an adult man looking at two familiar toys (one gender-stereotyped for boys, and one gender-stereotyped for girls). The experimenter introduced each story, and then the rest played as a pre-recorded video. The Toy Store trial involved a boy and his uncle buying a toy from the toy store (doll vs. truck, see Figure 8). The Carnival trial involved a boy winning a game at a fair and choosing a prize (pink bear vs. blue bear). Across participants, trial order and toy position were counterbalanced.

Note that both stories were always about a young boy and a male adult. While the underlying inferences here could well hold with gendered stereotypes about young girls (as we explore more in Experiment 3), we focused on boys because their gender counter-stereotypical

behaviors and preferences are typically policed more by adults than girls (e.g., Kane, 2006), and thus we expected that the inference from speaker surprisal would be most likely.

Each video showed a brief conversation. In both conditions, the target boy initially requested a toy (e.g., “Can we get a toy for my birthday?”) and the adult acknowledged and accepted the request fluently (e.g., “Yeah, let’s get one of those toys for your birthday”). This initial back-and-forth was included to establish that the child is allowed to choose a toy, and to demonstrate that the adult sometimes responds fluently. Next, the target boy requested one of the toys (e.g., “I want that one please”). Critically, the target boy’s selection was ambiguous from the participant’s perspective, as there was no visual cue to indicate which toy the child selected.

Test. In both conditions, the test utterances were positive and affirming of the character’s choice. In the fluent baseline condition, the adult responded fluently (e.g., “Sure, honey. We can buy you that one”). In the surprise condition, the adult responded with the same permissive message but with markers of surprise and production difficulty (e.g., “Oh really? Um... Sure, honey. Uh... We can buy you that one”). Participants were then asked which toy the target boy asked for (our primary dependent measure).



Figure 8. A still from Experiment 1 showing the basic experimental setup for the Toy Store trial (toy position counterbalanced).

Results

Guided by our preregistered analysis plan, we first tested for sensitivity to feedback, with separate regressions predicting toy choice from condition for each age group. We see a significant effect of condition on 6- to 7-year-old children's responses ($\beta = 0.26, p = .011$) and 8- to 9-year-old children's responses ($\beta = 0.26, p = .007$). These condition effects showed that older children were selecting the "girl" stereotyped toy more frequently in the surprise condition (see Figure 9). There was no effect of condition on 4- to 5-year-old children's responses ($\beta = -0.01, p = .949$). Note also that, unsurprisingly, children in the fluent baseline showed significant gender stereotypes, predicting boys would select a "boy" stereotyped toy in all three age groups ($ps < .001$).

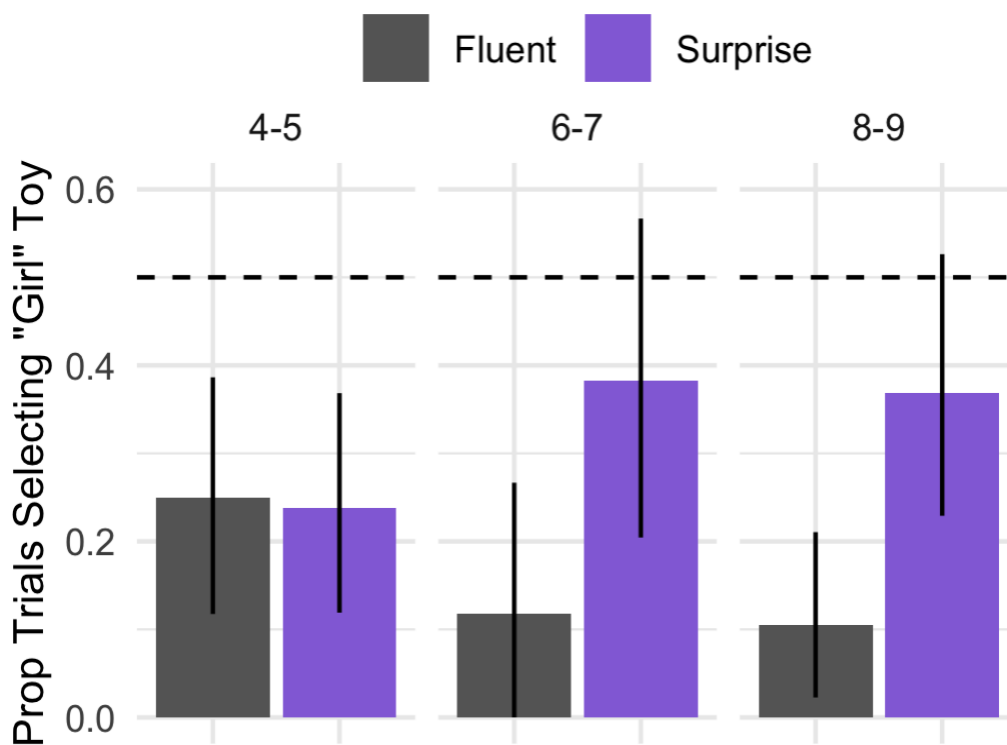


Figure 9. Children's toy selections across conditions for each of our three pre-determined age bins for Experiment 1. Error bars show bootstrapped 95% confidence intervals.

Discussion

We find that by age 6-to-7 children are more likely to infer that a boy chose a counter-stereotypical toy (e.g., a doll) if an adult responds with surprisal markers, compared to baseline. While children at all ages showed clear gender stereotypes at baseline, older children were able to partly override this stereotype based on an adult's surprisal. These data provide an initial demonstration that children are connecting conversational cues of surprisal with expectations about gender stereotypes. Thus, even though the parent gave a permissive and egalitarian response, when their linguistic markers revealed that they seemed surprised, 6-to-7 year-old children were relatively more likely to assume counter gender-normative behavior.

Experiment 2

In Experiment 2, we use a novel alien environment to ask whether these surprisal cues can provide a possible learning mechanism for developing new expectations about normative behavior. While Experiment 1 demonstrates that children connect surprisal cues with extant beliefs about other's expectations, this may or may not implicate these cues in the learning of new expectations (i.e. forming a new stereotype may be more complicated than linking a reaction to an established stereotype). Thus, Experiment 2 directly tests whether conversational surprisal cues can enable *learning* a novel expectation.

Rather than relying on pre-existing gender stereotypes to inform participants' priors about what is expected, Experiment 2 used novel behaviors and categories (aliens called "Hibbles" wearing hats). By manipulating surprisal cues, we aimed to differentially establish the exact same novel behavior as either unmarked and equally expected (fluent baseline) or marked and potentially unexpected (surprise condition). To do so, our primary measure asked

participants to directly evaluate the markedness of the target behavior (judging it as normal or weird), rather than inferring the behavior that evoked surprise (as in Experiment 1).

Method

Participants

We collected data from a pre-registered sample of 120 children ages 4-to-9, with 20 children in each condition in each of three pre-registered age bins (4-5, 6-7, 8-9). This experiment was being conducted remotely over Zoom, with a live experimenter present. As with Experiment 1, children completed this experiment over Zoom, interacting with a live experimenter who navigated a slide-style, animated Qualtrics survey. Based on pre-registered exclusion criteria, an additional 5 children were excluded and replaced due to technical difficulties, failing to answer all the questions, or parent interference.

A separate sample of 80 adults were recruited via Amazon Mechanical Turk and paid \$0.75 for their participation. Adult participants completed the same task, but adults navigated the task on their own via Qualtrics. Participants who failed a CAPTCHA or a simple auditory attention check were prescreened and unable to complete the study.

Procedure

Participants were shown an animated story that the experimenter narrated. Participants were introduced to a novel alien group (“Hibbles”) and told about a school with a Hibble teacher and three Hibble students getting ready for a party. The rest of the story played out in a pre-recorded video wherein each Hibble child put on a hat one-at-a-time and the Hibble teacher responded affirmatively to each one. Each Hibble put on a different colored hat (red, green, and yellow, with colors counterbalanced across participants, see Figure 10 for a visualization of the

task). The three response utterances followed the same structure with some variation (i.e. varying the initial response token across the three utterances “nice”, “yeah”, “cool”).

In both conditions, the pattern of choices was identical and the teacher responded fluently to the first two Hibbles’ hats. Across conditions, we manipulated the teacher’s response to the third Hibble’s choice (hereafter referred to as the target). In the fluent baseline condition, the teacher responded fluently, comparable to the past selections (e.g., “Cool. You look great!”). In the surprise condition, the teacher responded with stilted surprise, while still affirming the choice as before (e.g., “Oh! Um... Cool. You look uh... great!”).

As our primary measure, participants were then asked to evaluate the normality of the target’s choice (“Do you think it’s normal or weird for a Hibble to wear a [green] hat?”, with a two-point contingent follow-up question, e.g., “a little [weird] or really [weird]?”). As follow-up measures, participants were also asked to predict what color hat a novel Hibble would wear (prediction measure), and told about a Hibble who had been teased and asked to infer which color hat that the Hibble *had been* wearing (teasing measure).

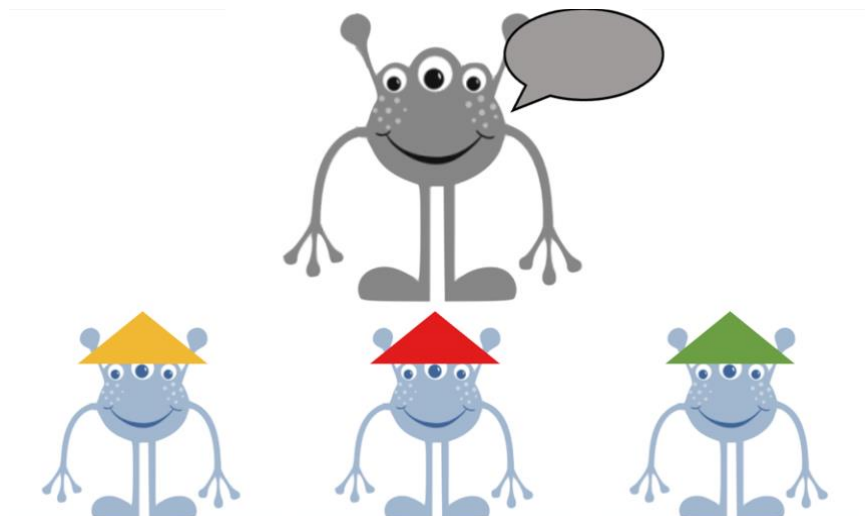


Figure 10. A still from Experiment 2 showing the Hibbles and their hats (colors counterbalanced).

Results

Following our pre-registered analysis plan, for each measure, we first report overall regression models, testing for the effects of condition, age (measured continuously), and their interaction, and then follow-up analyses testing the effect of condition in each predetermined age bin.

For children’s weirdness judgments (our primary measure, see Figure 11), we see a significant effect of condition ($\beta = 0.48, p = .013$) such that children judged the target behavior as weirder in the surprise condition, and marginal interaction effect between condition and age ($\beta = 0.23, p = .050$). Examining children’s weirdness judgements separately for each age bin, we see a significant effect of condition with the 8- to 9-year-olds ($\beta = 0.85, p = .008$), a marginal effect with the 6- to 7-year-olds ($\beta = 0.60, p = .052$), and no effect with the 4- to 5-year-olds ($\beta = -0.06, p = .894$). That is, older children, but not younger, judged that wearing the target hat color was weirder when it had elicited a surprisal reaction, compared with the fluent condition.

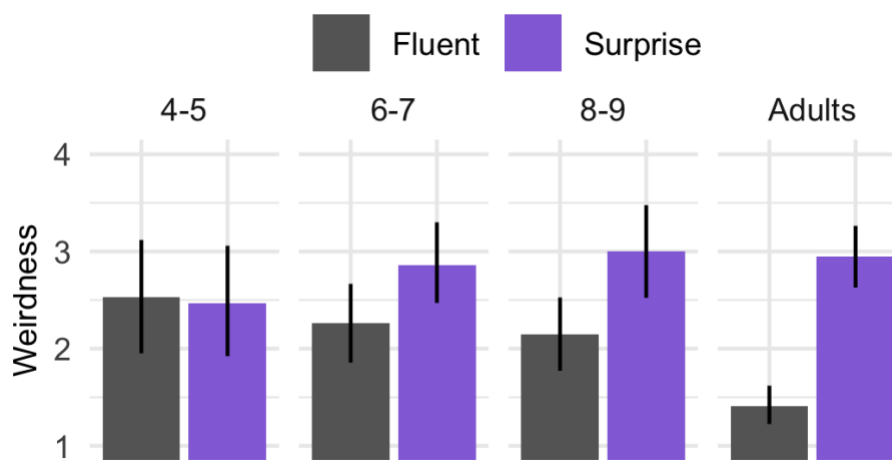


Figure 11. Children’s weirdness judgements across conditions for each of our three pre-determined age bins for Experiment 2, with the adult sample for comparison. Error bars show bootstrapped 95% confidence intervals.

We next turn to our two follow-up measures. Examining children's predictions, we found no significant effects of condition ($\beta = 0.10, p = .199$), age ($\beta = -0.01, p = .853$), or their interaction ($\beta = 0.03, p = .566$). Examining children's responses for the teasing measure, we found a significant effect of condition ($\beta = 0.40, p < .001$) such that children were more likely to expect that a teased character had been wearing the target hat color in the surprise condition, and no significant effect of age ($\beta = 0.04, p = .244$) or their interaction ($\beta = 0.02, p = .639$). When asked about a novel Hibble who was teased, children in every age group were more likely to infer that Hibble had been wearing the target hat color in the surprise condition, compared with the fluent condition (see Figure 12, all $ps < 0.05$).

Adult Results

For adults, we see significant effects of condition for all our measures. Adults in the surprise condition judged the target hat color as significantly weirder, compared with adults in the fluent baseline condition (see Figure 11; $\beta = 1.54, p < 0.001$). Adults in the surprise condition were less likely to predict that a new Hibble would wear the target hat color, compared with adults in the fluent baseline ($\beta = -0.19, p = 0.04$). Adults in the surprise condition were also more likely to expect that a Hibble who was teased had been wearing the target hat color, relative to the fluent baseline (see Figure 12; $\beta = 0.44, p < 0.001$).

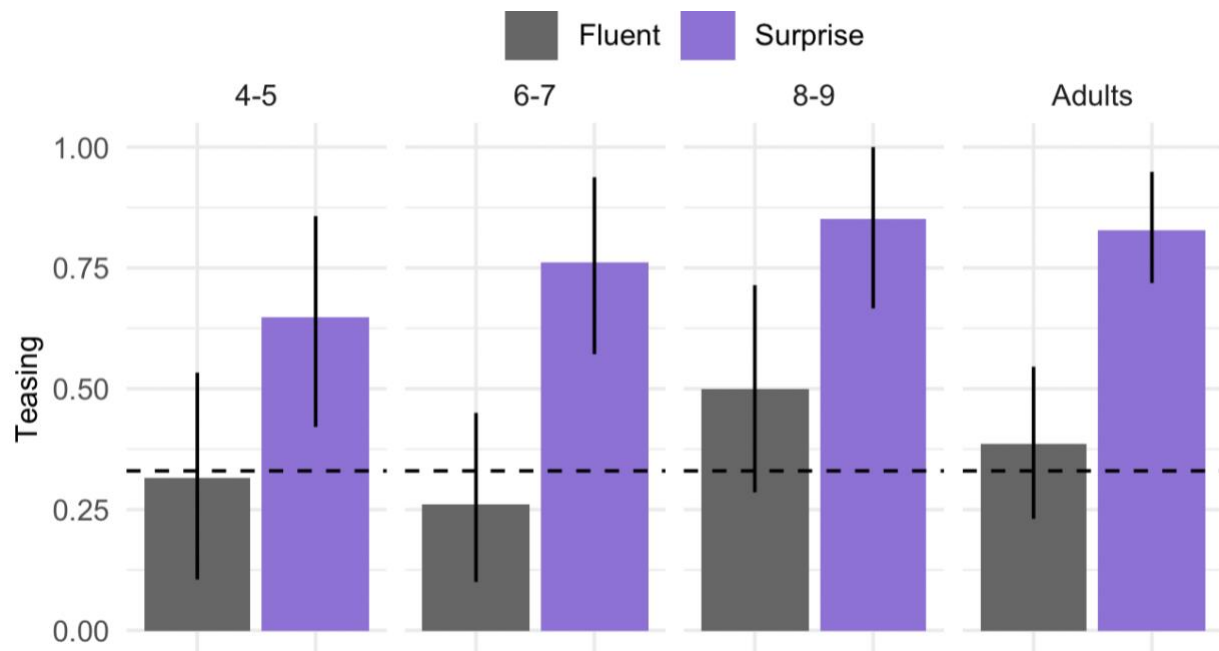


Figure 12. Children’s responses for the teasing measure across conditions for each of our three pre-determined age bins for Experiment 2, with the adult sample for comparison. Higher numbers indicating selecting the target hat color as the cause of teasing (dashed line indicates chance with three color options). Error bars show bootstrapped 95% confidence intervals.

Discussion

Experiment 2 demonstrates that conversational cues to surprisal may serve as a viable learning mechanism for transmitting novel speaker expectations, and potentially stereotypes. Adults readily use other’s surprisal reactions to learn a novel expectation, generate predictions, and infer social consequences. The developmental data clearly show that older children are sensitive to the feedback type in their weirdness evaluations (our primary measure), while 4–5-year-old children do not show any sensitivity to feedback (as in Experiment 1). For children’s predictions about a novel Hibble, we saw no effect of feedback type which could suggest children are not incorporating surprise into their own predictions, although null effects are

difficult to interpret and there might have been difficulties detecting this effect with our choice measure (i.e. a reduction in selections against a 33% chance baseline). Interestingly for the teasing measure, children at all ages in the surprise condition inferred that a character was teased for wearing the target hat, more so than the fluent condition. Overall, these results suggest that surprise cues license additional inferences not just about extant expectations (as in Experiment 1), but also for learning entirely new and consequential expectations.

Experiment 3

Experiment 3 returns to the domain of gender stereotypes to ask how adults might use surprisal cues to learn a novel, gendered expectation. We introduced participants to a novel kids game called “Blickets” and show some students who are playing Blickets (always an equivalent number of boys and girls). Unlike the prior experiments, Experiment 3 also contrasts two surprisal conditions to further probe the flexibility of adults’ inferences. In one surprisal condition, the surprisal reactions covary with gender (gendered-surprise condition), while in the other they happen for both boys and girls (control-surprise condition). As before, we also contrast these two surprisal conditions with a fluent baseline condition.

We predicted that adults would incorporate information about both the presence and distribution of surprisal feedback when drawing inferences. We again used a perceived weirdness measure to capture unexpectedness, and predicted both surprisal conditions would lead to perceived unexpectedness relative to baseline. We also included two measures probing the extent to which adults saw the game as gendered, and predicted that only the gendered-surprise condition would stand out on those measures, and not the control-surprise condition (where surprise may be attributed to something more idiosyncratic).

Method

Participants

A pre-registered sample of 150 adults (50 per condition) were recruited via Prolific and paid \$0.80 for their participation. Participants who failed a CAPTCHA or a simple auditory attention check were prescreened and unable to complete the study.

Procedure

Participants were randomly assigned to one of three conditions: fluent baseline condition, gendered-surprise condition, or a control-surprise condition. Participants read a short animated story about a classroom where some of the kids like to play a game called “Blickets”. In the story, four children (two boys and two girls) come to the teacher one at a time to ask for a toy to play Blickets. After each child asks for a toy, participants heard pre-recorded audio of the teacher’s response for that child, affirming their choice. The four response utterances followed the same structure with some variation (i.e. varying the initial response token across the four utterances “nice”, “yeah”, “cool”, “sure”).

Across conditions, we varied the surprisal of the teacher’s responses. In the fluent baseline condition, the teacher provided unmarked responses to all children (e.g., “Yeah, you can play Blickets.”). In the gendered-surprise condition, the teacher provided fluent responses for two students of one gender, but used surprisal markers for two students of the other gender (e.g., “Oh! Um... Yeah, uh... you can play Blickets.”). In the control-surprise condition, the teacher also provided surprisal responses for two students, but now for one boy and one girl. From hereon, we will refer to the last child was the “target” (and this child received the exact same surprisal response in the two surprise conditions). Across participants, we counterbalanced the order of the children with two orders varying the final target’s gender: boy-target order (girl,

boy, girl, boy) and girl-target order (boy, girl, boy, girl). Please refer to Figure 13 for a simplified schematic of each condition.

Participants were then asked 3 dependent measures in a fixed order (using 7-point bipolar scales, with 4 indicating neutrality). For the weirdness measure, participants were asked to judge if the teacher thought it was normal or weird that the target character wanted to play “Blickets” (1 - really weird to 7 - really normal). For the teasing measure, participants saw two novel characters (a boy and a girl) who also played “Blickets” and were asked to predict which had been teased (1 - probably Bryan to 7 - probably Olivia). Lastly for the stereotype measure, participants were asked who usually plays “Blickets” (1 - mostly boys to 7 - mostly girls). Note that for analysis purposes, we reverse coded the teasing and stereotype scales for the girl-target-order, so that we could compare responses across orders.












Condition					Inference
Fluent Baseline	✓	✓	✓	✓	
Gendered Surprise	✓		✓		
Control Surprise		✓	✓		

Figure 13. A schematic showing the logic for each of the three conditions for Experiment 3 (check marks indicate a fluent reaction, surprise icons indicate a surprisal reaction). Note, this schematic shows only the boy-target order for simplicity.

Results

First for weirdness judgments, adults inferred the teacher thought the target’s behavior was weirder in both the gendered-surprise ($\beta = -3.66, p < 0.001$) and control-surprise conditions ($\beta = -2.81, p < 0.001$), relative to the fluent baseline (see Figure 14). Comparing our two surprisal conditions, adults inferred the teacher thought the target’s behavior was significantly weirder in the gendered-surprise condition ($\beta = -0.85, p < 0.01$), compared with the control-surprise condition.

For teasing predictions (see Figure 14), adults were more likely to infer the target’s gender was teased in the gendered surprise condition relative to the control-surprise condition ($\beta = -1.78, p < 0.001$) and fluent baseline ($\beta = -1.67, p < 0.001$), as predicted. Similarly, adults were also more likely to infer that the game was gendered in the gendered-surprise condition relative to the control surprise condition ($\beta = 1.81, p < 0.001$) and fluent baseline ($\beta = 2.09, p < 0.001$). They did not differentiate the control surprise and fluent baseline conditions on either measure (all $ps > 0.22$).

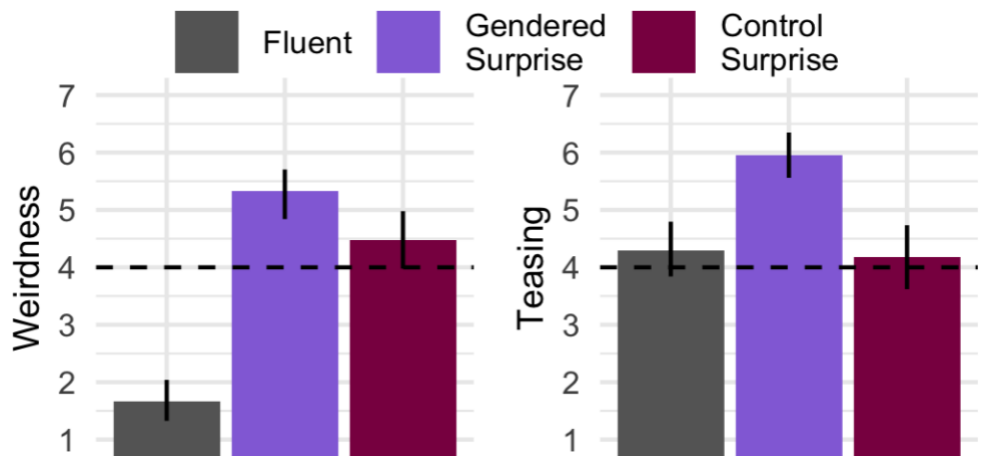


Figure 14. Adults’ judgements for the weirdness (left) and teasing (right) measures for each of the three conditions in Experiment 3 (note we did not collect developmental data for this experiment). For teasing, higher values indicate selecting the character who was same gender as the target. Error bars show bootstrapped 95% confidence intervals.

Discussion

Experiment 3 demonstrates that adults readily integrate surprisal information and statistical covariance. After hearing a surprisal reaction (in both surprisal conditions), adults rated the target's behavior as weirder in the teacher's eyes, compared with the fluent baseline. However it was only when those surprisal reactions covaried with gender (gendered-surprise) that adults inferred that the novel game was gendered and also used gender to infer who was teased. Although there was an equal amount of surprise in control-surprise condition, adults did not infer that the game was gendered or use gender to infer who was teased. Interestingly, adults also rated that the teacher thought the target behavior was weirder in the gendered-surprise condition than the control-surprise condition (despite hearing the exact same audio clips), which may be further evidence that they are inferring a possible norm in the gendered-surprise condition.

General Discussion

In Chapter 2, across 3 experiments, we see consistent evidence that even well-intentioned feedback about a child's behavior can nonetheless reveal one's underlying expectations. In each experiment, even though the feedback across conditions was closely matched, the presence of markers of surprise and production difficulty (interjections "oh" and disfluencies "um") was sufficient to generate differentiated inferences. Experiment 1 demonstrates that children by age 6 to 7 use conversational markers of others' surprise to reason about whether a boy made a stereotypical or counter-stereotypical choice. Experiment 2 demonstrates that adults and older children use these same cues to learn a novel expectation and predict social consequences. Experiment 3 combines these approaches to show that adults use others' conversational surprise

to learn a novel gendered expectation. This work contributes to the recent “Emotion as Information” framework that argues emotional expressions are useful not just for reasoning about emotions, but for learning unobservable states in the social and physical world (Wu et al., 2021).

While these surprisal inferences clearly reflect reasoning about the speaker’s expectations, we remain agnostic as to whether they are seen as capturing descriptive or prescriptive information. Either way, these cues could serve as one mechanism for transmission of social stereotypes. Stereotypes and others’ expectations powerfully shape children’s own behavior and beliefs from early on in life. In the first year of life, children begin making gender-stereotyped toy choices (Todd, Barry, & Thommessenn, 2016). By the preschool years, children predict others’ preferences based on gender-stereotypes (Eisenberg, Murray, & Hite, 1982). Broadly speaking, the messages that children hear in conversation provide an important avenue for the transmission of pernicious social beliefs (Rhodes et al., 2012; Chestnut et al., 2021), but beyond isolated messages, the current research suggests that conversational feedback may also provide important information. Specifically, Chapter 2 provides novel experimental evidence of one potential mechanism of transmission: conversational cues that convey expectations. We have focused on gender stereotypes as a pernicious and naturalistic test case of stereotyped expectations; however, our proposal applies to learning a variety of expectations and stereotypes.

Our work adds new insights to the literature on belief transmission that demonstrates the surprising efficacy of subtle linguistic framing (e.g., Chestnut et al., 2021; Cimpian et al., 2007; Rhodes et al., 2012). Specifically, the current work shows that it is not just what we say, but how we say it that matters. The types of interjections and disfluencies we investigated here are highly naturalistic, paralinguistic features of casual language use that can convey information about a

speakers' mental states to young learners. Building off of our results from Chapter 1, an exciting question for future research is the extent to which children are reasoning about other's expectations here by beginning to model the production process that generated these speech cues. Alternatively, it is possible participants could be reasoning about these cues more heuristically, or even relying on other inferences about a speaker's underlying discomfort or dishonesty (Fox Tree, 2002).

We also note that while our work has focused on interjections and disfluencies as conversational markers of surprisal or production difficulty, there are likely a number of conversational reactions that would spark similar inferences. Indeed, our account should generalize well to any response that indexes the speaker's expectations or their (dis)preference. For example, other responses that operate similarly might include indicating uncertainty ("Are you sure?"), spotlighting the preferred choice ("What about this one?"), otherwise signaling cost and unwillingness ("Well, if that's the one you really want, I guess..."), and more. In exploring the generality of these kinds of cues, it will also be interesting to explore how conversational markers interact with the previously examined role of non-verbal affect (Skinner et al., 2020). For example, future work could test how surprisal interjections ("oh") that vary in affect show overlapping inferences (e.g., about an underlying expectation) or differentiable inferences (e.g., about the implications of that surprise).

Across Experiments 1 and 2, the data suggest 4- to 5-year-olds are not reliably using others' surprisal to draw inferences. While even infants connect surprisal reactions with expectations about the physical world (Wu et al., 2024), it is possible that younger children in our experiments struggle to connect their representation of the adult's expectations with an additional representation of others' behaviors and mental states. We also note that the

developmental pattern we observe is consistent with related work on reasoning about an agent's competence on the basis of others' facial expressions of surprise—tasks that likely have similar representational complexity (Asaba et al., 2020). However, it is also possible that younger children can draw the key inference, but their performance is burdened by task demands.

In continuing this line of research, it will be important to explore the extent to which these cues are available in children's naturalistic language environments (e.g., in conversations with caregivers). For example, parents and children could be asked to discuss wordless picture books displaying characters making gender stereotypical and gender counter-stereotypical choices. We could then code caregiver behavior to look for conversational cues to surprisal, production difficulty, and discomfort in the counter-stereotypical trials. Specifically, it would be interesting to look for interjections (e.g., "oh!"), disfluencies, surprised facial expressions, and more. Having these kinds of data in hand would further flesh out a story wherein these cues are (1) a potential mechanism for communicating expectations, (2) reflecting and potentially perpetuating stereotyped beliefs, and (3) readily available in children's own environments. We could then potentially show these videos to new participants and see what they glean from these naturalistic interactions.

Conversations carry a wealth of social information, especially conveying a speaker's underlying beliefs (e.g., Rhodes et al., 2012). Even well-meaning or explicitly egalitarian messages can sometimes still carry pernicious social messages (Chestnut et al., 2021). Children burgeoning abilities to extract underlying belief information from language helps them learn about the social world very quickly, which might be unfortunate in cases where adults are inadvertently conveying stereotype information.

Chapter 3: Listener design and listener knowledge

In Chapter 1, children evaluated the mental implications of a speaker's pauses, judging their knowledge and preferences based on *how* (and how quickly) something was said. In Chapter 2, children reasoned about the mental states underlying how someone *reacts*, reasoning about others' surprise and learning what is expected. In Chapter 3, we explore how conversation can prompt inferences about someone, even before that person has spoken, based on how they are spoken *to*. To illustrate, imagine joining a friend for a drink and meeting a mutual friend (Sam) for the first time. You mention that you recently saw a Scorsese movie, and your friend turns to Sam and says "Martin Scorsese is a director who did *The Departed* and a bunch of other films." Before Sam has even spoken, you're able to form an expectation about their underlying knowledge in this case—namely that Sam is likely unfamiliar with Scorsese. Chapter 3 explores this ability and its development in early childhood.

The inference above is guided by the principle that language is meant to be *informative*. But that informativity is fundamentally contextual—one must be informative given our current interlocutor's existing knowledge. As a result, how we speak is shaped by our assumptions about an audience's knowledge (i.e. audience design or listener design, Clark & Murphy, 1982). Smooth conversation necessitates knowing what kind of knowledge can be assumed (and thus go unexplained) and what kind of knowledge is at issue (and needs to be put forth). A large body of work has documented the variety of ways in which adults tailor their communication, calibrating the amount and kind of information they provide by taking the audience's knowledge into account (e.g., Clark & Murphy, 1982; Brown-Schmidt & Hanna, 2011). For example as speakers, adults reduce the amount of information they give when re-telling a story to someone who has heard it before, but not when re-telling the story to a new partner (Galati & Brennan,

2010). As listeners, eye-tracking studies demonstrate that adults also readily account for speaker knowledge when interpreting an utterance, such as distinguishing shared and privileged knowledge (Hanna et al., 2003). While there has been some debate about whether language starts from a more egocentric default (e.g., Horton & Keysar, 1996; Keysar et al., 2000), it is clear that adults routinely engage in listener design under many circumstances. Caregivers even readily engage in listener design with their children; in experimental demonstrations, caregivers show remarkably fine-grained and in-the-moment adjustments to their children’s vocabulary knowledge (Leung, Tunkel & Yurovsky, 2021). Given that listener design is based on what the listener already knows, just hearing how someone is spoken to may offer hints about that person’s knowledge, even without context. Indeed, this is exactly the kind of logic underlying why condescension feels so obnoxious. We return to these observations after discussing children’s ability to engage in listener design.

There is extensive evidence that even young children engage in listener design as well. While classic work suggested children struggle to adapt to their communicative partners even into middle childhood (e.g., Krauss & Glucksberg, 1977), such tasks may have involved undue cognitive load that masked the competence of younger children. In simplified behavioral tasks, children as young as 12-months adjust their referential gestures based on their interlocutor’s knowledge state—for example, gesturing to a hidden toy’s location more when their interlocutor did not witness the hiding (Liszkowski et al., 2008; O’Neill, 1996). Children by at least age 4-to-5 show even clearer hallmarks of listener design in their conversations. Shatz and Gelman (1973) demonstrated that children as young as age 4 adjust their language when talking with a 2-year-old interlocutor, compared with an older child interlocutor. By age 5, children show remarkably nuanced adjustments—giving general information about an object category (e.g., “cups are for

drinking stuff”) when talking to a naïve interlocutor, and specific information (e.g., “this cup is dotty”) when talking to a knowledgeable interlocutor (Baer & Friedman, 2018). Also by age 5, children are beginning to form an explicit understanding of these adjustments and make predictions on the basis of language register—e.g., expecting that so-called infant-directed speech will be directed to a baby (Labotka & Gelman, 2020). Learning language is about much more than acquiring a formal system, and this work clearly demonstrates that young children are developing a rich suite of skills that enable them to *use* language effectively to be understood by different audiences.

The above research indicates that adults and even children have impressive listener design abilities in productive and receptive language. Building on this foundation, the current work explores a novel way of testing children’s understanding of the relationship between language and listener knowledge. Listener design is predicated on the idea that you (the speaker) first have an understanding of the listener’s knowledge and then you can adjust your speech accordingly. If my friend knows nothing about X-men comics, I’ll have to explain who each character is as I bring them up, talking in a way that presupposes little common ground on the topic. In principle, this connection opens the possibility for the inverse inference— to also use aspects of how someone speaks to infer the *listener’s* underlying knowledge state. Even without context, seeing me talking with my friend, you’re likely to infer they know little about X-men or even comics in general based on how I talk. If smooth conversation necessitates knowing what kind of knowledge can be assumed (and thus go unexplained), then seeing what knowledge needs to be explained can imply what knowledge is (presumed to be) unshared.

The current work probes this hypothesized connection—first demonstrating that adults indeed reason in this way and then asking when these inferences arise in development. These

inferences likely require recruiting some understanding of the listener design process (albeit informally) to reason that a speaker produced their utterance based on the listener's knowledge. The ability to reason about others' mental states (or others' impressions of those mental states) from how someone is spoken to can provide important social information to young learners, and allow children to reason about condescension and other conversational subtleties that may rely on these skills.

Experiment 1a

As this inference process has not been well established in adults, we first begin with an experiment to establish that adults do indeed infer a listener's knowledge based just on how the listener is spoken *to*. We manipulated listener design between-subjects such that participants either saw speakers provide basic descriptions for familiar objects (labelling and offering general information), or non-basic descriptions (describing and offering specific information about each object) to their listener. The key question was not how these descriptions reflect speaker's knowledge, but instead how they might reflect the *listener's* likely knowledge, even before they have spoken. The main prediction was that participants would infer that someone who is *told* basic information about an object is likely less knowledgeable about that object. Note that this initial demonstration also allowed us to validate our manipulation of listener design description (basic vs. non-basic).

Method

Participants

60 participants were recruited from Amazon Mechanical Turk, with 30 participants in each of the two conditions (basic vs non-basic). Participants were paid a small reward in

exchange for completing the study. Participants who failed a CAPTCHA or a simple auditory attention check were prescreened and unable to complete the study.

Procedure

Participants read a short, animated story about young children. Across three trials, participants were shown brief conversational exchanges between two children at a time, one of whom (the speaker) described a familiar item to the other (the listener). Participants were randomly assigned to one of two listener design conditions manipulating the speaker's descriptions (basic vs. non-basic). In the basic description condition, the speaker used language designed for a naïve audience, unfamiliar with the target object and its kind, that presupposed little shared knowledge (e.g. "This is an astronaut. Astronauts get to go to space", more information on stimuli construction below). In the non-basic description condition, the speaker used language designed for a more neutral audience, presupposing some familiarity with the object's kind (e.g. "This is a cute astronaut. I like playing space with this astronaut"). Note that this experiment was conducted without pre-recorded audio stimuli, thus participants read through the animated story, including the target utterances. In each trial, participants were asked to evaluate the *listener's* knowledge ("How much do you think [this person] knows about astronauts?") on a scale from 1-to-7. Participants completed three trials with different characters and objects, and trial order was counterbalanced across participants.

In designing our basic description utterances, we used labelling statements (e.g., "This is an astronaut") and general facts true of the category and phrased with the generic syntax (e.g., "Astronauts get to go to space"). In designing our non-basic comparison stimuli, we adapted those utterances to instead use description (e.g., "This is a cute astronaut") and specific facts about this particular exemplar and the speaker's idiosyncratic preferences (e.g., "I like playing

space with it”). Note that the utterances were designed to be similar where possible while varying the amount of underlying audience knowledge they presuppose. These stimuli were designed partly by considering the language that children themselves produce for knowledgeable and ignorant audiences (Baer & Friedman, 2018).

Results

To test the effect of listener design, we ran a mixed effects linear regression predicting knowledge judgement by listener design condition, with a random effect of subject. As predicted, there was a significant effect of listener design condition ($\beta = 1.07, p = .009$), such that adults in the basic description condition rated the listener as less knowledgeable ($M = 2.83, SD = 1.80$), compared with adults in the non-basic description condition ($mean = 3.90, SD = 1.25$).

Discussion

This first experiment provides proof of concept that adults can readily reason about someone’s knowledge based on how that person is spoken to. When someone was told more basic, categorical information about an object, adults inferred that person was relatively less knowledgeable about the domain. These data suggest that adults may have some intuitive understanding of listener design, at least reasoning that speakers provide information that is new given the listener’s current knowledge.

Experiment 1b

Experiment 1a established that adults can use listener design in order to make inferences about the listener’s knowledge; even before a person has spoken, people inferred that someone who had something basic explained to them was less knowledgeable than someone who heard a more neutral explanation. Next, we adapted this task in a pre-registered experiment with children

ages 4- to 9-years-old to see when in development children infer knowledge just from how someone is spoken to.

We made two key adjustments to simplify the task for young children, based on our adult findings in Experiment 1a. First, rather than a between-subjects manipulation asking children to judge each listener individually, we contrasted two different descriptions side-by-side and asked children to select a listener. That is, children were shown a speaker who talks to two different listeners about a familiar object— offering one a basic description of the object (labelling and offering general information), and the other a non-basic description (describing and offering specific information about that object). This contrast was intended to simplify the task and children’s response (using a two-alternative forced choice measure), and to highlight the adjustment phenomenon. Secondly, given that adults inferences were largely driven by an inference of ignorance in the basic description condition, we also simplified our dependent measure to be a question focused on relative ignorance (described below).

Method

Participants

We collected data from a pre-registered sample size of 60 children, with 20 children in each of three pre-registered age bins: 20 4- to 5-year-olds (mean age = 5.02, 12 girls), 20 6- to 7-year-olds (mean age = 6.93, 11 girls), and 20 8- to 9-year-olds (mean age = 8.93, 10 girls). This experiment was conducted remotely over Zoom, with a live experimenter present. Participating families were recruited via a participant database largely of Chicagoland families who have previously participated in in-person research studies.

Procedure

Children were shown an animated story that an experimenter narrated (see Figure 15). Children were introduced to a target character (Jane) who had brought some toys into her class at school. Children were shown a toy that Jane had brought (e.g., an astronaut toy), and told that Jane knew that one of her classmates had never seen a toy like this one before. We included this background to ensure that children understood Jane was acquainted with the listeners relevant knowledge states. Jane then showed the toy to two different listeners one-at-a-time. Jane gave one listener a basic description (e.g. “This is an astronaut. Astronauts get to go to space”), and the other a non-basic description (e.g. “This is a cute astronaut. I like playing space with this astronaut”). The two listeners were given color-coded labels—the “blue girl” and the “green girl”—to facilitate children’s ability to respond verbally to this task (as these data were collected remotely on Zoom where verbal responses as these are much easier to reliably capture—e.g., versus pointing). Children were asked to make an ignorance judgement (“Which person has never seen a toy like this before— the [blue] girl or the [green] girl?”). Children completed two trials about the same people discussing different familiar toys (an astronaut and a submarine). Across participants, trial order and which listener heard the low-knowledge description first were counterbalanced. The three target characters remained the same across trials, and were always girls. After completing the final trial, children were asked to offer an open-ended explanation (“Why do you think the [green] girl has never seen a toy like this one before?”).

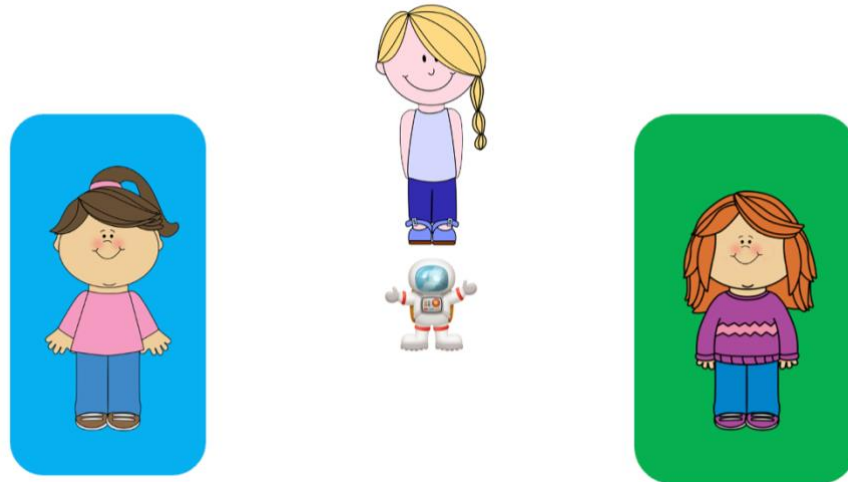


Figure 15. Example stimuli showing the basic trial structure layout for Experiment 1b.

Results

Our pre-registered analysis examined speaker selections separately for each of our three age groups (see Figure 16). 6- to 7-year-old (proportion of trials = 0.71, $p < 0.01$) and 8- to 9-year-olds (proportion of trials = 0.97, $p < 0.001$) reliably selected the basic listener as ignorant about the toy, whereas 4- to 5-year-olds did not (proportion of trials = 0.56, $p = 0.51$). In other words, older children inferred that someone is less knowledgeable when that person was told a basic level description, compared with the non-basic.

Discussion

We found that by age 6, children can readily make inferences about a listener, even before they have spoken, based on how they were spoken to— inferring that someone who hears a basic knowledge description (that implies low listener knowledge) is more likely ignorant about the subject at hand. That is, they seem to have inferred that the speaker was engaging in listener design and used this to make inferences about the limits of that listener’s knowledge—if this listener knew what the word astronaut meant, the speaker would not have needed to explain what it was. We see evidence of developmental change in this ability, and pre-registered age group

comparisons suggest that this reasoning was not present in 4- to 5-year-olds, but was present in older children.

This work presents an initial step in understanding how children use the way someone speaks not just to make inferences about the speaker, but also to make inferences about the listener. While a great deal of prior work has studied children’s knowledge attribution for speakers (e.g., Koenig et al., 2015), these data show how utterances can also imply things about the listener’s knowledge. While reasoning about the speaker and the listener likely share some common cognitive processes, it is interesting to consider whether reasoning about the listener requires any additional cognitive skills. One possibility is that reasoning about the listener may require some second order theory of mind skills to the extent that people see the speaker’s utterance as reflecting their *belief* about the listener’s knowledge. In Experiment 2, we aimed to explore the extent to which children in this task are reasoning about the speaker’s beliefs about the listener.

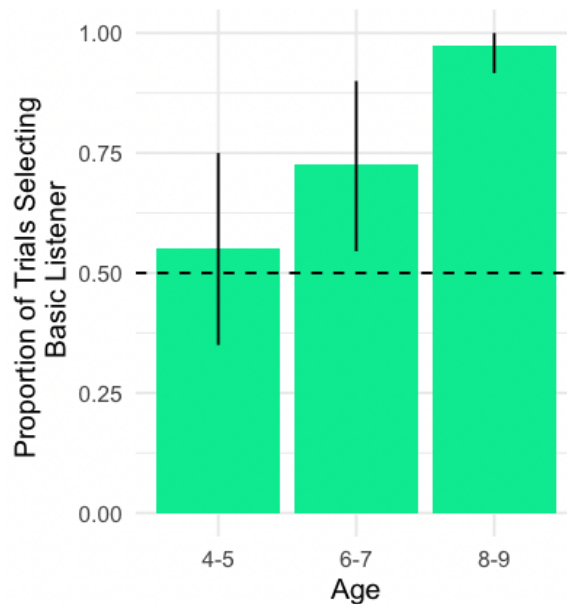


Figure 16. Age-binned results for Experiment 1b showing the proportion of trials where children selected the basic listener as ignorant. Dashed line indicates chance responding.

Experiment 2

Experiment 1b demonstrated that children are able to use listener design principles to infer listener knowledge based on how the listener is spoken to, even before they have spoken. Such an inference rests on the idea that the speaker is providing information to each listener calibrated to their knowledge. Of course, rather than a true signal of the listener's knowledge, a speaker's utterance only reflects their *assumptions* about the listener's knowledge, which may or may not match reality. Indeed, cases of condescension reflect a mismatch in exactly that assumption, with the speaker underestimating the listener's true knowledge (e.g., “mansplaining”). In Experiment 2, we investigated whether children understand that listener design reflects a speaker's *belief* and that those beliefs can sometimes be disconnected from the listener's actual knowledge. This experiment thus helps us understand how deeply children are reasoning about the process of listener design by testing cases where a speaker's appraisal of someone's knowledge seems to differ from ground truth.

In a pre-registered experiment with children ages 4 to 9-years-old, we manipulated listener design and background knowledge, while also asking questions that probe both the listener's knowledge and speaker's appraisal. Across trials, we manipulated listener design by having some speakers use basic language and some speakers use non-basic language when describing familiar objects. In addition to manipulating what the speakers say, we also manipulated the background knowledge of the listener: participants either received no information about each listener's familiarity with the target domain (no background) or were explicitly told that each listener is familiar with the target domain (background given). By including a condition where the listener should have knowledge (background given), the goal was to create a situation that might prompt participants to consider a discrepancy between the target's actual knowledge

(which should be high) and the speaker's appraisal (which might be relatively lower for the basic description trials).

Additionally, we also moved away from the two-alternative forced choice design employed in Experiment 1b to instead collect graded judgements of each listener, instead of providing a contrast and asking a relative measure (i.e. we asked "How much does this person know about [X]" rather than "Who do you think has never seen [X] before"). This provides a more difficult test of how spontaneously children use listener design to infer listener knowledge because each individual item only receives one description, rather than presenting children with two contrasting alternatives (as in Experiment 1b).

First for the knowledge measures, we predicted the background manipulation would shift children's knowledge judgements, consistent with prior work. Children in the background-given condition should overall judge the listener to be more knowledgeable than children in the no background condition. Additionally, we also expected that children's knowledge judgements in the no-background condition would follow our findings in Experiment 1b, with older children inferring the basic listener has less knowledge. For the appraisal measures, we predicted an effect of our listener design manipulation where hearing the teacher offer a basic description means the teacher thinks less of the listener's knowledge. This would mean that the teacher appraisal ratings are lower for the basic listener compared with the non-basic listener, regardless of the background manipulation.

The key prediction was that older children's knowledge judgements would be minimally affected by listener design in the background given condition (establishing listener knowledge) even when the speaker uses basic language. Such a finding would indicate older children are able to reason about the speaker's *appraisal* separately from the listener's underlying knowledge.

While it is possible that using basic language reduces lowers judgements of the listener's knowledge in the presence of independent cues to knowledge (in the background-given condition), the reduction should be much larger for the appraisal judgements of the teacher's beliefs. Statistically, this prediction would be borne out by an interaction between listener design and background condition predicting children's knowledge judgements knowledge judgements, and no such interaction for the appraisal judgements.

After the dissertation proposal, the scope of Experiment 2 was expanded to add additional individual difference measures capturing executive functions and second order theory of mind (and aimed to recruit a larger sample to be able to examine possible individual differences). As a result of that expansion, piloting was slowed and data collection remains ongoing, thus currently we will report a preliminary look at the current sample. Due to statistical power constraints for our secondary measures with the incomplete sample, we present initial findings only for our primary measures here, and not the additional individual difference measures.

Method

Participants

We pre-registered a sample size of 180 children, with 30 children in each condition for each of three pre-registered age bins (4-5, 6-7, 8-9). The sample reported here data from 84 children currently collected: 23 4- to 5-year-olds (mean age = 4.79, 12 girls), 29 6- to 7-year-olds (mean age = 6.90, 13 girls), and 32 8- to 9-year-olds (mean age = 9.12, 17 girls). This study is conducted remotely over Zoom, with a live experimenter present. Participating families were recruited via a participant database largely of Chicagoland families who have previously participated in in-person research studies.

Procedure

In the task, children are introduced to a school with many teachers and many students. In the familiarization phase, children are introduced to a visual 5-point circle scale from 1 - “not much” to 5 - “a huge amount”. As a basic check of the scale, children are asked to make knowledge evaluations about a student who incorrectly labels a familiar animal (calling a pig a “horse”), and a character who appropriately labels a dump truck. Then, during each trial, children observe a teacher talking with a student about a target object. To manipulate the target’s underlying knowledge, children are randomly assigned to one of two between-subjects background knowledge conditions (background given and no background). In background given condition, the story specifies that the student has familiarity with the target object and of the general domain (e.g., “[This student] really likes astronauts. She has lots of astronaut toys at home and loves to read books about astronauts”) and this is intended to imply strong underlying knowledge. In no background condition, there is no information provided about the student’s familiarity with the general domain. In the story, the teacher shows a toy item to the student and describe it using either basic or non-basic language (varying description type within-subjects across trials for different items). Children are asked to make two key evaluations: (1) a knowledge inference (“How much does [the student] know about [astronauts]”), and (2) an appraisal inference (“How much does the teacher think [the student] knows about [astronauts]?”). These questions are asked in a fixed order to facilitate children’s reasoning about the embedded clause (and embedded mental representation) in the appraisal inference. Future studies could address possible order effects, but importantly this fixed order is the same in all conditions and thus any systematically different responding cannot be accounted for by this fixed order. Children respond on to all measures with a visual 5-point circle scale from 1 - “not much” to 5 - “a huge amount”. Children complete 4 trials with different characters, objects, and

descriptions. In alternating trials, the teacher provides a basic description or a non-basic description (i.e., two of each following an ABAB format) and we counterbalance trial item order as well as which items are paired with which description type across participants.

Results

First as a measure check, we examine children's knowledge judgements on the scale familiarization trials and find that children's responses indicate they are able to use the scale appropriately. That is, children report higher domain knowledge for a character who is able to correctly label a familiar item ($M = 3.95$, $SD = 0.73$) compared with a speaker who incorrectly labels a familiar item ($M = 1.62$, $SD = 0.96$). Children's judgements for these two practice measures are significantly different in every age group (all $ps < 0.05$).

As a manipulation check, we next examine children's knowledge judgements across the two background conditions (background-given vs. no-background) and find that children's judgements show sensitivity to our background manipulation, as expected given past research using similar manipulations. Averaging across description types, children in the background-given condition overall judge the characters as being more knowledgeable ($M = 4.32$, $SD = 0.81$) about the relevant domain compared with children in the no-background condition ($M = 3.09$, $SD = 1.09$). Children's knowledge judgements show a significant effect of background condition consistent with this pattern in every age group (all $ps < 0.05$; see Figure 17).

To test the key effect of listener design on children's rating knowledge ratings (how listener design impacts children's evaluation of the listener), we ran a linear regression predicting knowledge ratings with a main effect of listener design description. Contrary to our predictions, we find no overall effect of listener design description (basic vs. non-basic) for the knowledge measures ($\beta = -0.09$, $p = .59$). It is possible younger children are masking the competence of

older children, so we also ran the above regression separately for our oldest age group focusing only on the participants in the no background condition, which provides the closest comparison to Experiment 1b. However, here too we see no significant effect of listener design condition on children's knowledge ratings ($\beta = -0.04, p = .90$), with older children rating the listener's knowledge similarly regardless of whether the listener was given a basic description ($M = 3.21, SD = 1.26$) or a non-basic description ($M = 3.18, SD = 1.06$).

We next examine children's responses on the appraisal ratings. We ran a linear regression predicting appraisal ratings with a main effect of listener design description. As with children's knowledge ratings, we find no overall effect of listener design description (basic vs. non-basic) for the appraisal measures ($\beta = 0.15, p = .42$). Looking separately at the oldest children's appraisal judgements, we still see no effect of listener design description ($\beta = 0.50, p = .12$), with older children rating the speaker's belief about the listener's knowledge similarly regardless of whether the listener was given a basic description ($M = 2.93, SD = 1.21$) or a non-basic description ($M = 3.43, SD = 1.14$).

As our initial prediction concerning the impact of speaker description was not borne out in these data, we are unable to ask how children may track a speaker's appraisal separately from the listener's underlying knowledge, as we had planned. We discuss these unexpected findings more below and discuss why we may have obtained a null here despite children's apparent sensitivity to listener design in Experiment 1b.

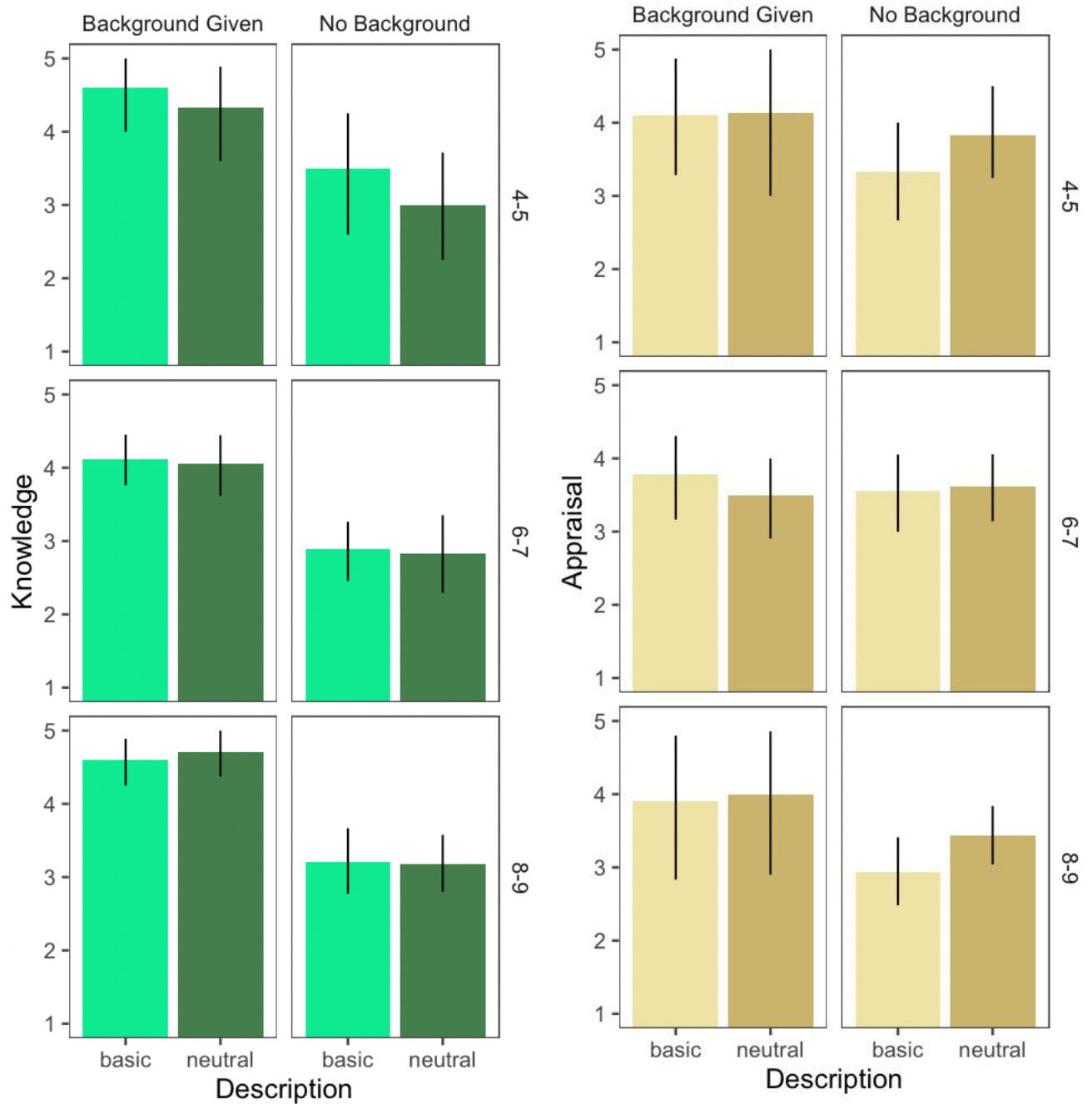


Figure 17. Data for Experiment 2 showing children’s judgements by description type for the knowledge measure (left panes) and appraisal measure (right panels) for each age bin, split by background condition.

Discussion

Experiment 2 was intended to replicate the effects of Experiment 1b in a more complex paradigm (i.e. no forced choice) and to explore the nature of the inferences the make about listener design by manipulating the listener's underlying familiarity with the topic. Unlike Experiment 1b, we found a null result of listener design here: that is children did not evaluate a listener who was told a basic description as less knowledgeable. While these developmental data are incomplete and thus remain preliminary, we have no evidence that even the oldest children in our sample are successfully making these inferences in this task.

Is this because they understood nothing about the task? We do not think so. We see that children are indeed able to use our knowledge scale appropriately when making simple speaker judgments (i.e., about someone who is correct or incorrect). Additionally, we also see that our manipulation of the listener's familiarity with the topic was successful in inducing children at all ages to expect a familiar listener was more knowledgeable than the baseline where no such information was provided.

The null effect of listener design observed in Experiment 2 is perhaps surprising given the strength of the effects we observe for in Experiment 1b for 6- to 7-year-olds and 8- to 9-year-olds. In Experiment 1b, 8- to 9-year-olds almost unanimously inferred that a character who was told a basic description was likely ignorant; yet their knowledge judgements in Experiment 2 show no corresponding difference by description type. We think there are several possible reasons for this null and so caution too strong of an interpretation of this null until we conduct follow up work.

It is worth noting that Experiment 2 tested a new response format wherein children were asked to judge listeners one-at-a-time. One possibility is that a more direct contrast between two possible descriptions for the same item is necessary for children to draw these listener inferences

(as was present in Experiment 1b). In Experiment 2, the children did not see the speaker address two different listeners, and seeing a speaker treat one person as relatively less knowledgeable might be key for children to make these inferences. In the absence of this more direct contrast, it is possible that generating a knowledge inference based on listener design requires generating a relevant alternative utterance that could have been said, and children may struggle to come up with alternatives spontaneously on their own. Indeed, at least younger children do seem to struggle with generating the relevant alternatives that are crucial for many kinds of pragmatic inferences (e.g., Barner et al., 2011). Although adults in Experiment 1a used speaker description to infer listener knowledge without any contrast (i.e. in a fully between-subjects design), the results of Experiment 2 suggest that children are not using speaker description to infer listener knowledge in this task, contrary to our predictions and the results of Experiment 1b.

In addition to removing the contrasting descriptions, there were also several other, smaller changes made to the task design across Experiments 1b and 2. For example, the use of a more targeted, ignorance-focused measure in Experiment 1b may have proved more sensitive for capturing this inference with children. Additionally, the initial stipulation in Experiment 1b that one of the characters may lack knowledge could also facilitated children's sensitivity to the implications of how someone explains. Lastly, it is possible that asking the knowledge measure together with the more linguistically and conceptually complex appraisal question ("How much does the teacher think [Sam] knows about astronauts") may have muddled children's responses overall in Experiment 2. It is difficult to interpret null effects thus hard to say which of these changes (or their combination) may have contributed to children's difficulties in Experiment 2. We return to this issue to discuss possible next steps in the General Discussion that follows.

Without an effect of description in Experiment 2, we are unable to demonstrate the key interaction we expected between the effect of speaker description and background. Nonetheless, future studies should aim to test the question of how children understand listener design as a reflection a speaker's *beliefs* separate from the listener's underlying knowledge in a more sensitive task, perhaps returning to the contrasted design from Experiment 1b. Understanding the developmental trajectory of distinguishing those inferences will provide important insights into how richly children are reasoning about the listener design process.

General Discussion

In conversation, how we speak is fundamentally sensitive to (what we think) our listener knows already, determining what can be assumed (and go unsaid) and what needs to be explained. Experiments 1a and 1b provide initial, novel evidence that adults and children reason about this process of listener design and are able to make the corresponding inverse inference: that someone who is told basic information about a topic might be unfamiliar with it. In these experiments, adults and children from age 6- to 7-years-old use a speaker's utterance to evaluate the listener's knowledge. However, in Experiment 2, where we attempted a more complex design that did not include forced choice measures, children did not distinguish between listener design conditions, which we discuss more below. While adults show these inferences robustly even in non-forced choice contexts, we see no evidence that children can do so. Still, when given a forced choice measure, older children recognize that someone who receives a more basic explanation likely knows less than someone who does not receive a basic explanation.

The null effect of listener design in Experiment 2 is difficult to interpret at this time. We are currently working on a task that more closely mirrors the design of Experiment 1b by returning to a more contrastive setup wherein children hear two possible descriptions for the same item—where a speaker addresses two different listeners. To the extent that such a contrast is indeed necessary, it raises the question of how naturalistic this kind of reasoning might be, since most conversations likely resemble the between-subjects setting. While seeing directly contrasted descriptions may be less common in everyday life, the school classroom may actually be one setting where one would often see similar topics explained differently, or additional explanation directed toward specific children, in ways that provide a natural contrast. In piloting our updated task, we are also more explicitly establishing possible knowledge variation (i.e. that some of the characters might know a lot, and others might not know much at all). Future studies will aim to clarify the relative robustness or fragility of children’s listener design inferences.

We have long known that a speaker’s utterance tells you a lot about how knowledgeable the speaker is, but this work demonstrates that such utterances can also imply things about a listener’s knowledge. It is interesting to note that prior work shows that children by age 5 expect generic claims to be more widely known (Cimpian & Scott, 2012), and also attribute greater knowledge to *speakers* who offer generically presented information, rather than specific (although children were dependent on verifiable information until age 7; Koenig et al., 2015). By asking about the *listener’s* knowledge in this work, we see that adults and older children make precisely the opposite inference about the listener based on somewhat similar utterances. This is interesting in part because it suggests that children may believe that the basic description is more informative (i.e. it implies the speaker has more knowledge), but precisely for that reason indicates that the *listener* actually knows less (and needed to be told). This interesting dynamic

highlights the new questions and nuances that come from embedding utterances in conversational settings and considering principles of how language is *used*, rather than just focusing on isolated, decontextualized utterances. Children are coming to understand and actively reason about the basic tenet of listener design—to meet our listener where they are at—and thus evaluate the implications of such a principle for the listener’s knowledge. In this way, this work also adds to the literature on listener design in children by providing a novel way of testing children’s understanding of the relationship between listener knowledge and language production.

At the heart of Chapter 3 is the idea that how someone talks to you carries rich information about how they are thinking about you. While it may sound abstract, such a reasoning process likely underlies a variety of everyday conversational experiences, such as condescension (i.e. a case where there is a mismatch between how you are being addressed and your underlying knowledge state). Broadly, Chapter 3 prompts new questions about how children themselves come to detect and respond to other’s appraisals of them as in cases like perceived condescension—topics which have received little empirical attention. Anecdotally, children seem to engage in a variety of spontaneous behaviors aimed at demonstrating knowledge their own in the face of possible slights (e.g., protest behaviors such as saying “duh” or “I already knew *that*”), but this topic has not been well examined in the literature. Cognitively, detecting and reacting to other’s knowledge appraisals is a complex feat that may recruit mental state reasoning, second-order theory of mind, reputation management, and other skills. Additionally, detecting others’ appraisals in this way (and protesting accordingly when misguided) may play an important and underrecognized role in social learning by helping us demarcate what we *already* know—allowing us to avoid unnecessary redundancy, while also

generally signaling competence (in a situation where perceived competence may be under threat). Where active learning and information-seeking allow children to fill gaps in their existing knowledge, these behaviors could help children readjust other's perceptions and clarify what is known already.

Conclusion

Chapter 3 provides proof-of-concept that we can use listener design to make knowledge inferences about the listener; how someone is spoken *to* can provide insight into their likely knowledge. Adults robustly reason about how language is used in this way, and infer that someone who is told more basic, category level information about a topic is likely less knowledgeable about that topic, even before they've spoken. When presented with contrasting descriptions (one basic and one neutral), children by age 6-to-7 infer that the listener who heard the more basic description is likely unfamiliar. More research is needed to establish the richness and limitations of children's ability to reason about listener design in this way. How someone speaks to you can signal their appraisal of your knowledge, and understanding that fact has implications for a range social competencies, such as condescension.

General Discussion

Across three chapters, we see a picture emerging of children (ages 4-to 9-years-old) developing abilities to derive social information from rich, contextual interpretations of subtle conversational cues. In Chapter 1, children use a speaker's fluency to reason about their mental states, separate from the content of what is said—e.g., even if both people answer correctly, children infer knowledge based on how easily those responses came. Even 4-to 5-year-old children use disfluency to infer a speaker's knowledge and show some flexibility in their reasoning, while older children show even more impressive flexibility to make distinctive inferences about knowledge and preferences depending on the conversational context. In Chapter 2, children extend this kind of thinking to reason about the implications of conversational cues in feedback—an adult's approving words can be understood very differently when accompanied by markers of surprise and production difficulty (“Oh! Um... Sure”). Children and adults use these markers to infer a speaker's unstated expectations, reason about pre-existing stereotyped beliefs, and even learn entirely novel social expectations. In Chapter 3, children and adults are even able to draw inferences about a listener before they have spoken, at least when descriptions are contrasted. Together, these three chapters demonstrate that *how* something is said has profound implications for the kind of inferences we draw about a speaker's underlying mental states. Testing a range of conversational cues—disfluencies, interjections, descriptions—and a breath of social inferences—about utterances, feedback, and listeners—this dissertation opens a new space of questions about how children come to understand speech as a reflection of thought.

We argue that children's inferences across these studies may best be characterized as a process of rational inference guided by an intuitive model of how mental states relate to communication. Taking Chapter 1 as an example, children not only understand the implications

of not being able to speak as quickly and easily as expected, but also flexibly understand the same cue (speaker disfluency) in different ways. Specifically, a disfluency preceding a correct answer leads to lower inferences of speaker knowledge, but a disfluency preceding an “I don’t know” causes no hit to the speaker’s knowledge. Older children even draw distinctive generalizations about the meaning of a disfluency depending on whether it’s coming from the labelling context or the preference context. Looking across such demonstrations, we argue children’s reasoning process is fundamentally inferential, rather than being explained by low-level features, simple behavioral heuristics, or general stereotypes. Together, the research in this dissertation provides an initial step towards understanding how children reason about the relationship between utterances and the mental states and processes that generate them.

Deriving mental information from language likely recruits a range of potentially distinct but interrelated skills and inferences. The three chapters in this dissertation provide different ways of assaying children’s developing ability to do so. The inferences at the heart of this dissertation are in no way exhaustive, and we remain agnostic about the extent to which they reflect shared or distinct developmental processes. That is, these tasks were specifically designed to capture (at least some of) the range and conceptual variation necessary for drawing mental inferences from subtle communicative cues. Nevertheless, it is interesting to note that across the chapters we see similar developmental patterns with children from age 6-to-7 are drawing flexible and robust inferences across our tasks, despite that variability in the cues and conceptual representations involved in each task. We do see some evidence that younger children ages 4-to-5 are able to draw some of the key mental inferences (as in Chapter 1); however, their ability to do so seems more limited, at least in the current tasks. As future research develops a more comprehensive account of how children understand language as a reflection of thought, it will be

interesting to examine the extent to which developmental data reflect more of a shared underlying infrastructure, or reflect distinctive skills with differing trajectories over development. We have attempted to capture coarse versions of this with some of the tasks we used in Chapter 3, but ideally we would develop a battery to potentially capture the important skills that underlie competencies in the kinds of tasks we have explored here (see Bohn et al., 2023 for such an examination of pragmatic development).

In this dissertation, we've seen how a speaker's subtle conversational cues (disfluencies, interjections, listener design) can prompt inferences about a speaker's knowledge, preferences, expectations, and even their listener's knowledge. As noted above, this list is not exhaustive and likely only scratches the surface of describing the breadth of questions to investigate. For example, our perspective here could also be fruitfully applied to understand how children make use of subtle listener feedback in real time during conversation (e.g., backchannels, facial expression, etc.) to track understanding and adjust their own behavior—e.g., introducing new information and realizing you've lost someone before they say so (e.g., Bacso et al., 2021; Bacso & Nilsen 2022). In adults, these micro-calibrating sequences are thought to be crucial to the grounding process (i.e. achieving mutual understanding) and their systematicity and complexity have been well attested in largely descriptive work (Bavelas & Gerwing, 2011; Bavelas et al., 2017). In most any snapshot of conversation, there is dynamic mental information leaking out in real time and speakers and listeners must quickly and jointly make sense of this information and its implications.

While it is clear that children and adults readily infer others' mental states from how they speak, it is important to note that we are not making any claims about the veracity of these inferences. While some work with adults suggests that, for example, listener's perceptions of a

speaker's knowledge are tied to the same cues speakers actually produce when less knowledgeable (Brennan & Williams, 1995; Smith & Clark, 1993), such a relationship does not need to be the case for listeners to draw inferences that are systematic, robust, and rational. As a parallel, decades of demonstrations in cognitive psychology emphasize that actual human behavior is rife with irrationalities (Kahneman, 2011); however, our commonsense psychology (i.e. how we reason about others' actions) does seem to be well captured by a process of reasoning about others as rational agents (Jara-Ettinger et al., 2016). Similarly, it is likely that children and adults' inferences in this dissertation reflect reasoning over an idealized, rational model of speech production, rather than a veridical model per se.

What's in others' minds? Conversations, emotions, and actions: A broader framework

While this dissertation probes children's social reasoning in specifically conversational contexts, it is crucial to also consider and situate this work within the context of children's mental state reasoning in domains beyond language and communication. By toddlerhood, children are adept at reasoning about other's goals from their actions (e.g., Woodward, 1998) and desires from their expressions (e.g., Repacholi & Gopnik, 1997). Across the preschool years, children are able to more explicitly reason about other's mental states, demonstrating a "Theory of Mind" and mastering the so-called false belief task (e.g., Gopnik & Wellman, 1994; Wellman, 2014). False belief understanding represents the gold standard for demonstrating theory of mind because these tasks require children to represent others as holding inaccurate beliefs about the true state of the world (e.g., predicting an agent will search an empty box if that was the last place they saw a target toy). Children also integrate false belief into their language-based reasoning, for example, accounting for a speaker's false belief in interpreting their claim (e.g., Robinson & Mitchell, 1992), and understanding that language can correct false beliefs (e.g.,

Song et al., 2008). It is worth noting that false belief represents just one aspect of epistemic reasoning and this overly narrowed focus has limited our ability to create a comprehensive account of children's epistemic reasoning abilities; more recent theoretical accounts and empirical work in mental state reasoning has begun to push beyond this traditional Theory of Mind framework (e.g., Baker et al., 2017; Jara-Ettinger et al., 2016; Rubio-Fernandez, 2018).

While acknowledging this past work, we note that the vast majority of research on children's (and adult's) mental state reasoning has focused on physical actions. Work in language and communication has received comparatively little attention. Recently scholars have noted this unfortunate oversight, arguing that much could be gained from a greater focus on conversational mental state reasoning, and specifically that more dialogue is needed between research on Theory of Mind and research on experimental pragmatics (Rubio-Fernandez, 2018; Westra & Negel, 2021). This dissertation also adds to these calls by highlighting conversation as the central venue for exercising mental reasoning skills in everyday life, even in early childhood.

Conversational mental state reasoning is a crucial area of study not only because conversation is likely the most ubiquitous context for mental state reasoning, but also because social reasoning is so fundamental to language understanding (Goodman & Frank, 2016). One key area that has begun to explore some of these key connections is work on experimental pragmatics which includes a large body of phenomena and inferences—from scalar implicature to hyperbole to irony—and that breadth (and divergent findings across phenomena) could inspire novel thinking in the Theory of Mind literature (which has largely focused on false belief understanding). Indeed, some recent empirical research has begun to draw these connections between pragmatics and theory of mind more clearly in work with adults (Jara-Ettinger & Rubio-Fernandez, 2021), and children (Jara-Ettinger et al., 2020). For language especially, much could

be gained from focusing on children’s epistemic reasoning in everyday conversational contexts where children must make interesting communicative inferences that might not require false belief representations but instead require other abilities to make deep inferences about others’ minds (Bohn & Frank, 2019; Rubio-Fernandez, 2018). Everyday mental state reasoning in conversation likely involves a host of other inferences that a comprehensive account of the development of mental state reasoning would need to capture.

Our account is inspired by a set of recent interrelated frameworks proposed for language (Goodman & Frank, 2016), social learning (Gweon, 2021), commonsense psychology (Jara-Ettinger et al., 2016), and emotional expressions (Wu et al., 2021). The Naïve Utility Calculus is one recent proposal of a unified, formal account that argues mental state reasoning is built on an understanding that agents select actions by maximizing rewards and minimizing cost (Jara-Ettinger et al., 2016). This parsimonious account can capture a wide swath of commonsense psychology, from action predictions about how an agent is expected to behave given their mental state (e.g., being willing to incur a higher cost when desire is high) to inverse inferences about the mental state that likely gave rise to an observed behavior (e.g., that someone who paid a high cost must highly value the reward). Recent empirical studies suggest that 10-month-old infants already demonstrate mental state reasoning captured by naïve utility calculus—expecting an agent to prefer a reward they were willing to incur a higher cost for (Liu et al., 2017). Preschool-age children are able to make a variety of inferences consistent with naïve utility calculus about an agent’s preferences (Jara-Ettinger et al., 2015), desires (Jara-Ettinger et al., 2017), and knowledge (Aboody et al., 2021). While this model has largely been developed for and applied to reasoning about physical actions, it provides a fruitful basis for considering what such a model

could look like to capture children's language-based inferences in this dissertation (see also Jara-Ettinger et al., 2019).

In considering what this intuitive model could look like, it may also help to consider extant frameworks in pragmatics for how speakers' utterances relate to their goals, the context, and their listener's possible interpretations. The Rational Speech Act framework (RSA; Goodman & Frank, 2016) provides one such account for formalizing pragmatic inferences. Very generally, RSA is a probabilistic model of language production and understanding via a process of recursive social reasoning that integrates information about communicative goals, shared common ground, and the space of things one could say. For example, to what extent are the fluency-based inferences of Chapter 1 related to pragmatic reasoning in implicature tasks? RSA (and accounts of pragmatic reasoning more generally) appeals heavily to the role of alternative utterances (e.g., the speaker could've chosen to say *Y*, but they chose to say *X* here instead), but it is unclear how heavily alternatives feature in fluency-based inferences (e.g., to capture graded inferences by timing; Roberts et al., 2011). While alterations would need to be made, such a model could perhaps account for fluency-based inferences with the inclusion of a speech cost parameter or incorporating a notion of time. The RSA framework has been successfully applied to a variety of phenomena from contextual reference (Frank & Goodman, 2012), vagueness (Lassiter & Goodman, 2017), hyperbole (Kao et al., 2014), and much more. Perhaps most relevantly to the ideas in this dissertation, it was recently expanded to account for politeness by imposing tradeoffs between competing speaker goals—to be informative, to be kind, and to present oneself well (Yoon et al., 2020). While these alterations are more substantial, they allow the model to capture richer subtleties of uniquely social language use that would be necessary to capture the inferences in this dissertation. Broadly, this dissertation provides new evidence for

considering the development of children’s pragmatic abilities, and demonstrates that much more could still be gained by focusing on the ways that our mental states leak out during our communication with others.

In developing an account of how we reason about mental states leaking out in conversation, it will be crucial to distinguish between reasoning about others’ mental *states* and others’ mental *processing*. Research on mental reasoning has focused almost exclusively on reasoning about mental states (e.g., beliefs and desires), however recent research has begun to explore how we are able to reason about agent’s ongoing mental *processes* (e.g., being lost in thought vs. thinking about a solution; Berke et al., 2023; Richardson & Keil, 2022). In many situations, the same mental state (e.g., knowing an answer) might be meaningfully different depending on the mental process that generated that state (e.g., having memorized that answer versus solving the problem in real time; Richardson & Keil, 2022). Recent proposals argue that we reason about others’ mental *processing* via a principle of rational mental effort (an extension of the framework provided by the Naïve Utility Calculus; Berke et al., 2023). Adults and older children seem able to infer other’s mental processes in this way in the context of problem solving (Berke et al., 2023; Richardson & Keil, 2022). Interestingly, while inferences of mental *states* are typically tied to actions, inferences of mental *processes* (at least in these demonstrations) seem instead to be crucially tied to the timing and pauses of those actions (Berke et al., 2023).

While this recent research focuses on physical problem solving, reasoning about other’s mental processes is especially ubiquitous and relevant during conversation. In conversation (as with physical actions), the timing and pauses in someone’s utterance prove critical to reasoning about their underlying processes. When the friend we are chatting with seems to trail off, we seem flexibly understand whether they’re trying to find their words or became distracted from

the conversation at hand. Indeed, the starts and stops in what someone says can signal their mental processes, from whether they are stuck trying to retrieve a word, remembering what they did last weekend, considering what's being asked of them, or planning how to politely decline (Brennan & Williams, 1995; Fox Tree, 2002; Roberts et al., 2011). While explicitly demarcating reasoning about mental *processes* rather than mental states is outside the current scope, this dissertation provides a novel testing ground for understanding how children reason about other people's thinking by paying attention to the subtleties of what they say.

Conclusion

Listening not just to what someone says, but also *how* they say it gives us a real-time glimpse into their underlying mental states and processes. In this dissertation, children demonstrate rich and flexible capacities to learn about their social world by paying attention to conversational cues in their environment. Across a variety of inferences, children are going well beyond the content of what is said and instead seem to be reasoning about production process behind the words. Children understand the mentalistic implications behind a speaker's pauses (Chapter 1), surprisal (Chapter 2), and over-explanations (Chapter 3). In every turn of conversation, we get a window into a speaker's thinking and processing, and together, the three chapters of this dissertation provide new insights into children as budding conversationalists and mentalists. Leveraging these mental reasoning skills will not only help young children become skilled conversationalists, but also facilitate their ability to engage in many of humanity's most striking cognitive feats—from social learning to cooperation.

References

- Aboody, R., Zhou, C., & Jara-Ettinger, J. (2021). In pursuit of knowledge: Preschoolers expect agents to weigh information gain and information cost when deciding whether to explore. *Child Development, 92*(5), 1919-1931.
- Akhtar, N., Carpenter, M., & Tomasello, M. (1996). The role of discourse novelty in early word learning. *Child development, 67*(2), 635-645.
- Arnold, J. E., Fagnano, M., & Tanenhaus, M. K. (2003). Disfluencies signal thee, um, new information. *Journal of psycholinguistic research, 32*, 25-36.
- Asaba, M., Wu, Y., Carrillo, B., & Gweon, H. (2020). You're surprised at her success? Inferring competence from emotional responses to performance outcomes. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 2650-2656). Austin, TX: Cognitive Science Society.
- Bacso, S. A., & Nilsen, E. S. (2022). Children's use of verbal and nonverbal feedback during communicative repair: Associations with executive functioning and emotion knowledge. *Cognitive Development, 63*, 101199.
- Bacso, S. A., Nilsen, E. S., & Silva, J. (2021). How to turn that frown upside down: Children make use of a listener's facial cues to detect and (attempt to) repair miscommunication. *Journal of Experimental Child Psychology, 207*, 105097.
- Baer, C., & Friedman, O. (2018). Fitting the message to the listener: Children selectively mention general and specific facts. *Child Development, 89*(2), 461-475.
- Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour, 1*(4), 0064.

- Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition*, *118*(1), 84-93.
- Barr, D. J. (2001). Trouble in mind: Paralinguistic indices of effort and uncertainty in communication. *Oralité et gestualité: Interactions et comportements multimodaux dans la communication*, 597-600.
- Barr, D. J. (2003). Paralinguistic correlates of conceptual structure. *Psychonomic Bulletin & Review*, *10*, 462-467.
- Bavelas, J. B., & Gerwing, J. (2011). The listener as addressee in face-to-face dialogue. *International Journal of Listening*, *25*(3), 178-198.
- Bavelas, J., Gerwing, J., & Healing, S. (2017). Doing mutual understanding. Calibrating with micro-sequences in face-to-face dialogue. *Journal of Pragmatics*, *121*, 91-112.
- Behne, T., Carpenter, M., & Tomasello, M. (2005). One-year-olds comprehend the communicative intentions behind gestures in a hiding game. *Developmental science*, *8*(6), 492-499.
- Bergey, C. A., & Yurovsky, D. (2023). Using contrastive inferences to learn about new words and categories. *Cognition*, *241*, 105597.
- Berke, M., Tenenbaum, A., Sterling, B., & Jara-Ettinger, J. (2023, May 8). Thinking about Thinking as Rational Computation. <https://doi.org/10.31234/osf.io/e65p3>
- Bernard, S., Proust, J., & Clément, F. (2014). The medium helps the message: Early sensitivity to auditory fluency in children's endorsement of statements. *Frontiers in Psychology*, *5*, 1412.
- Bian, L., Leslie, S. J., & Cimpian, A. (2017). Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science*, *355*(6323), 389-391.

- Birch, S. A., Akmal, N., & Frampton, K. L. (2010). Two-year-olds are vigilant of others' non-verbal cues to credibility. *Developmental science, 13*(2), 363-369.
- Bohn, M., & Frank, M. C. (2019). The pervasive role of pragmatics in early language. *Annual Review of Developmental Psychology, 1*, 223-249.
- Bohn, M., Schmidt, L. S., Schulze, C., Frank, M. C., & Tessler, M. H. (2022). Modeling individual differences in children's information integration during pragmatic word learning. *Open Mind, 6*, 311-326.
- Brennan, S. E., & Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of memory and language, 34*(3), 383-398.
- Brosseau-Liard, P. E., & Poulin-Dubois, D. (2014). Sensitivity to confidence cues increases during the second year of life. *Infancy, 19*(5), 461-475.
- Brown-Schmidt, S., & Hanna, J. E. (2011). Talking in another person's shoes: Incremental perspective-taking in language processing. *Dialogue & Discourse, 2*(1), 11-33.
- Bolinger, D. (1964). Around the edge of language: Intonation. *Harvard educational review, 34*(2), 282-296.
- Casillas, M. (2014). Taking the floor on time. *Language in interaction: studies in honor of Eve V. Clark, 12*.
- Casillas, M., & Frank, M. C. (2017). The development of children's ability to track and predict turn structure in conversation. *Journal of memory and language, 92*, 234-253.
- Chestnut, E. K., Zhang, M. Y., & Markman, E. M. (2021). "Just as good": Learning gender stereotypes from attempts to counteract them. *Developmental psychology, 57*(1), 114-125.

- Chomsky, N. (1980). A review of BF Skinner's Verbal Behavior. *The Language and Thought Series*, 48-64.
- Cimpian, A., Arce, H. M. C., Markman, E. M., & Dweck, C. S. (2007). Subtle linguistic cues affect children's motivation. *Psychological science*, 18(4), 314-316.
- Cimpian, A., & Scott, R. M. (2012). Children expect generic knowledge to be widely shared. *Cognition*, 123(3), 419-433.
- Clark, E. V. (2009). *First language acquisition*. Cambridge University Press.
- Clark, H. H. (1996). *Using language*. Cambridge University Press.
- Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, 84(1), 73-111.
- Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. In *Advances in psychology* (Vol. 9, pp. 287-299). North-Holland.
- Colomer, M., & Sebastian-Galles, N. (2020). Language background shapes third-party communication expectations in 14-month-old infants. *Cognition*, 202, 104292.
- Eisenberg, N., Murray, E., & Hite, T. (1982). Children's reasoning regarding sex-typed toy choices. *Child Development*, 81-86.
- Fox Tree, J. E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of memory and language*, 34(6), 709-738.
- Fox Tree, J. E. (2001). Listeners' uses of um and uh in speech comprehension. *Memory & cognition*, 29, 320-326.
- Fox Tree, J. E. (2002). Interpreting pauses and ums at turn exchanges. *Discourse processes*, 34(1), 37-55.

- Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084), 998-998.
- Freeman, N. K. (2007). Preschoolers' perceptions of gender appropriate toys and their parents' beliefs about genderized behaviors: Miscommunication, mixed messages, or hidden truths?. *Early childhood education journal*, 34, 357-366.
- Galati, A., & Brennan, S. E. (2010). Attenuating information in spoken communication: For the speaker, or for the addressee?. *Journal of Memory and Language*, 62(1), 35-51.
- Ganea, P. A., & Saylor, M. M. (2013). Talking about the near and dear: Infants' comprehension of displaced speech. *Developmental Psychology*, 49(7), 1299.
- Gelman, S. A., & Heyman, G. D. (1999). Carrot-eaters and creature-believers: The effects of lexicalization on children's inferences about social categories. *Psychological Science*, 10(6), 489-493.
- Gelman, S. A., Taylor, M. G., Nguyen, S. P., Leaper, C., & Bigler, R. S. (2004). Mother-child conversations about gender: Understanding the acquisition of essentialist beliefs. *Monographs of the society for research in child development*, i-142.
- Gelman, S. A., Ware, E. A., & Kleinberg, F. (2010). Effects of generic language on category content and structure. *Cognitive psychology*, 61(3), 273-301.
- Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 257-293). Cambridge University Press.
- Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41-58).
- Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11), 818-829.

- Gweon, H. (2021). Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in Cognitive Sciences*, 25(10), 896-910.
- Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, 49(1), 43-61.
- Harris, P. L., Koenig, M. A., Corriveau, K. H., & Jaswal, V. K. (2018). Cognitive foundations of learning from testimony. *Annual Review of Psychology*, 69, 251-273.
- Henderson, A. M., & Scott, J. C. (2015). She called that thing a mido, but should you call it a mido too? Linguistic experience influences infants' expectations of conventionality. *Frontiers in Psychology*, 6, 134065.
- Hilbrink, E. E., Gattis, M., & Levinson, S. C. (2015). Early developmental changes in the timing of turn-taking: a longitudinal study of mother–infant interaction. *Frontiers in psychology*, 6, 127539.
- Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground?. *Cognition*, 59(1), 91-117.
- Jara-Ettinger, J., Floyd, S., Huey, H., Tenenbaum, J. B., & Schulz, L. E. (2020). Social pragmatics: Preschoolers rely on commonsense psychology to resolve referential underspecification. *Child development*, 91(4), 1135-1149.
- Jara-Ettinger, J., Floyd, S., Tenenbaum, J. B., & Schulz, L. E. (2017). Children understand that agents maximize expected utilities. *Journal of Experimental Psychology: General*, 146(11), 1574.

- Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences*, 20(8), 589-604.
- Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2015). Children's understanding of the costs and rewards underlying rational action. *Cognition*, 140, 14-23.
- Jara-Ettinger, J., & Rubio-Fernandez, P. (2021). Quantitative mental state attributions in language understanding. *Science advances*, 7(47), eabj0970.
- Jaswal, V. K., & Malone, L. S. (2007). Turning believers into skeptics: 3-year-olds' sensitivity to cues to speaker credibility. *Journal of Cognition and Development*, 8(3), 263-283.
- Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.
- Kampa, A., & Papafragou, A. (2020). Four-year-olds incorporate speaker knowledge into pragmatic inferences. *Developmental science*, 23(3), e12920.
- Kane, E. W. (2006). "No way my boys are going to be like that!" Parents' responses to children's gender nonconformity. *Gender & Society*, 20(2), 149-176.
- Kao, J., Bergen, L., & Goodman, N. (2014). Formalizing the pragmatics of metaphor understanding. In *Proceedings of the annual meeting of the Cognitive Science Society* (Vol. 36, No. 36).
- Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, 11(1), 32-38.
- Kidd, C., White, K. S., & Aslin, R. N. (2011). Toddlers use speech disfluencies to predict speakers' referential intentions. *Developmental science*, 14(4), 925-934.
- Kinzler, K. D., Dupoux, E., & Spelke, E. S. (2007). The native language of social cognition. *Proceedings of the National Academy of Sciences*, 104(30), 12577-12580.

- Kinzler, K. D., Corriveau, K. H., & Harris, P. L. (2011). Children's selective trust in native-accented speakers. *Developmental science, 14*(1), 106-111.
- Kinzler, K. D., & DeJesus, J. M. (2013). Northern= smart and Southern= nice: The development of accent attitudes in the United States. *Quarterly Journal of Experimental Psychology, 66*(6), 1146-1158.
- Koenig, M. A., Cole, C. A., Meyer, M., Ridge, K. E., Kushnir, T., & Gelman, S. A. (2015). Reasoning about knowledge: Children's evaluations of generality and verifiability. *Cognitive psychology, 83*, 22-39.
- Koenig, M. A., & Echols, C. H. (2003). Infants' understanding of false labeling events: The referential roles of words and the speakers who use them. *Cognition, 87*(3), 179-208.
- Koenig, M. A., & Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child development, 76*(6), 1261-1277.
- Krauss, R. M., & Glucksberg, S. (1977). Social and nonsocial speech. *Scientific American, 236*(2), 100-105.
- Krehm, M., Onishi, K. H., & Vouloumanos, A. (2014). I see your point: Infants under 12 months understand that pointing is communicative. *Journal of Cognition and Development, 15*(4), 527-538.
- Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension?. *Language, cognition and neuroscience, 31*(1), 32-59.
- Kushnir, T., & Koenig, M. A. (2017). What I don't know won't hurt you: The relation between professed ignorance and later knowledge claims. *Developmental psychology, 53*(5), 826.
- Labotka, D., & Gelman, S. A. (2020). The development of children's identification of foreigner talk. *Developmental Psychology, 56*(9), 1657-1670.

- Lane, J. D., Wellman, H. M., & Gelman, S. A. (2013). Informants' traits weigh heavily in young children's trust in testimony and in their epistemic inferences. *Child development, 84*(4), 1253-1268.
- Lassiter, D., & Goodman, N. D. (2017). Adjectival vagueness in a Bayesian model of interpretation. *Synthese, 194*, 3801-3836.
- Leung, A., Tunkel, A., & Yurovsky, D. (2021). Parents fine-tune their speech to children's vocabulary knowledge. *Psychological Science, 32*(7), 975-984.
- Liberman, Z., & Shaw, A. (2020). Even his friend said he's bad: Children think personal alliances bias gossip. *Cognition, 204*, 104376.
- Liszkowski, U., Carpenter, M., & Tomasello, M. (2007). Pointing out new news, old news, and absent referents at 12 months of age. *Developmental science, 10*(2), F1-F7.
- Liszkowski, U., Carpenter, M., & Tomasello, M. (2008). Twelve-month-olds communicate helpfully and appropriately for knowledgeable and ignorant partners. *Cognition, 108*(3), 732-739.
- Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science, 358*(6366), 1038-1041.
- Martin, A., Onishi, K. H., & Vouloumanos, A. (2012). Understanding the abstract role of speech in communication at 12 months. *Cognition, 123*(1), 50-60.
- Meyer, W. U., Reisenzein, R., & Schützwohl, A. (1997). Toward a process analysis of emotions: The case of surprise. *Motivation and Emotion, 21*, 251-274.
- Moty, K., & Rhodes, M. (2021). The unintended consequences of the things we say: What generic statements communicate to children about unmentioned categories. *Psychological Science, 32*(2), 189-203.

- O'Neill, D. K. (1996). Two-year-old children's sensitivity to a parent's knowledge state when making requests. *Child development*, 67(2), 659-677.
- Orena, A. J., & White, K. S. (2015). I forget what that's called! Children's online processing of disfluencies depends on speaker knowledge. *Child development*, 86(6), 1701-1709.
- Pitts, C. E., Onishi, K. H., & Vouloumanos, A. (2015). Who can communicate with whom? Language experience affects infants' evaluation of others as monolingual or multilingual. *Cognition*, 134, 185-192.
- Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, 78(2), 165-188.
- Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: evidence from 14- and 18-month-olds. *Developmental psychology*, 33(1), 12.
- Rhodes, M., Leslie, S. J., & Tworek, C. M. (2012). Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences*, 109(34), 13526-13531.
- Richardson, E., & Keil, F. C. (2022). Thinking takes time: Children use agents' response times to infer the source, quality, and complexity of their knowledge. *Cognition*, 224, 105073.
- Roberts, F., Francis, A. L., & Morgan, M. (2006). The interaction of inter-turn silence with prosodic cues in listener perceptions of "trouble" in conversation. *Speech communication*, 48(9), 1079-1093.
- Roberts, F., Margutti, P., & Takano, S. (2011). Judgments concerning the valence of inter-turn silence across speakers of American English, Italian, and Japanese. *Discourse Processes*, 48(5), 331-354.
- Robinson, E. J., & Mitchell, P. (1992). Children's interpretation of messages from a speaker with a false belief. *Child Development*, 63(3), 639-652.

- Ruble, D. N., Martin, C. L., & Berenbaum, S. A. (2006). Gender Development. In N. Eisenberg, W. Damon, & R. M. Lerner (Eds.), *Handbook of child psychology: Social, emotional, and personality development* (pp. 858–932). John Wiley & Sons, Inc..
- Rubio-Fernandez, P. (2019). What Theory of Mind can learn from experimental pragmatics. In Cummins, C. & Kastos, N. (Eds.), *Handbook of Experimental Semantics & Pragmatics* (pp. 524-536). Oxford University Press.
- Sabbagh, M. A., & Baldwin, D. A. (2001). Learning words from knowledgeable versus ignorant speakers: Links between preschoolers' theory of mind and semantic development. *Child development, 72*(4), 1054-1070.
- Sabbagh, M. A., & Shafman, D. (2009). How children block learning from ignorant speakers. *Cognition, 112*(3), 415-422.
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1978). A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction* (pp. 7-55). Academic Press.
- Schützwohl, A., & Reisenzein, R. (1999). Children's and adults' reactions to a schema-discrepant event: A developmental analysis of surprise. *International Journal of Behavioral Development, 23*(1), 37-62.
- Shatz, M., & Gelman, R. (1973). The development of communication skills: Modifications in the speech of young children as a function of listener. *Monographs of the society for research in child development, 1-38*.
- Shriberg, E. (1996, October). Disfluencies in switchboard. In *Proceedings of international conference on spoken language processing* (Vol. 96, No. 1, pp. 11-14). Philadelphia, PA: IEEE.

- Skinner, A. L., Olson, K. R., & Meltzoff, A. N. (2020). Acquiring group bias: Observing other people's nonverbal signals can create social group biases. *Journal of personality and social psychology*, *119*(4), 824.
- Snow, C. E. (1972). Mothers' speech to children learning language. *Child development*, 549-565.
- Sobel, D. M., & Kushnir, T. (2013). Knowledge matters: how children evaluate the reliability of testimony as a process of rational inference. *Psychological Review*, *120*(4), 779.
- Soderstrom, M., & Morgan, J. L. (2007). Twenty-two-month-olds discriminate fluent from disfluent adult-directed speech. *Developmental Science*, *10*(5), 641-653.
- Song, H. J., Onishi, K. H., Baillargeon, R., & Fisher, C. (2008). Can an agent's false belief be corrected by an appropriate communication? Psychological reasoning in 18-month-old infants. *Cognition*, *109*(3), 295-315.
- Smith, V. L., & Clark, H. H. (1993). On the course of answering questions. *Journal of memory and language*, *32*(1), 25-38.
- Sperber D, Wilson D. 1995. *Relevance: Communication and Cognition*. Oxford, UK: Blackwell Publishers. 2nd ed.
- Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., ... & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*(26), 10587-10592.
- Todd, B. K., Barry, J. A., & Thommessen, S. A. (2017). Preferences for 'gender-typed' toys in boys and girls aged 9 to 32 months. *Infant and child development*, *26*(3), e1986.
- Tomasello M. 2008. *Origins of Human Communication*. Cambridge, MA: MIT Press.

- VanderBorgh, M., & Jaswal, V. K. (2009). Who knows best? Preschoolers sometimes prefer child informants over adult informants. *Infant and Child Development: An International Journal of Research and Practice*, 18(1), 61-71.
- Vouloumanos, A., Martin, A., & Onishi, K. H. (2014). Do 6-month-olds understand that speech can communicate?. *Developmental Science*, 17(6), 872-879.
- Vouloumanos, A., Onishi, K. H., & Pogue, A. (2012). Twelve-month-old infants recognize that speech can communicate unobservable intentions. *Proceedings of the National Academy of Sciences*, 109(32), 12933-12937.
- Westra, E., & Nagel, J. (2021). Mindreading in conversation. *Cognition*, 210, 104618.
- Wellman, H. M. (2014). *Making minds: How theory of mind develops*. Oxford University Press.
- White, K. S., Nilsen, E. S., Deglint, T., & Silva, J. (2020). That's thee, uuh blicket! How does disfluency affect children's word learning?. *First Language*, 40(1), 3-20.
- Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, 69(1), 1-34.
- Wu, Y., Baker, C. L., Tenenbaum, J. B., & Schulz, L. E. (2018). Rational inference of beliefs and desires from emotional expressions. *Cognitive science*, 42(3), 850-884.
- Wu, Y., & Gweon, H. (2021). Preschool-aged children jointly consider others' emotional expressions and prior knowledge to decide when to explore. *Child Development*, 92(3), 862-870.
- Wu, Y., Merrick, M., & Gweon, H. (2024). Expecting the Unexpected: Infants Use Others' Surprise to Revise Their Own Expectations. *Open Mind*, 8, 67-83.
- Wu, Y., Schulz, L. E., Frank, M. C., & Gweon, H. (2021). Emotion as information in early social learning. *Current Directions in Psychological Science*, 30(6), 468-475.

- Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2020). Polite speech emerges from competing social goals. *Open Mind, 4*, 71-87.
- Yoon, S. O., Jin, K. S., Brown-Schmidt, S., & Fisher, C. L. (2021). What's new to you? Preschoolers' partner-specific online processing of disfluency. *Frontiers in psychology, 11*, 612601.
- Yoon, S. O., & Fisher, C. (2020). Children's attribution of disfluency to different sources. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
- Zettersten, M. (2019). Learning by predicting: How predictive processing informs language development. *Patterns in Language and Linguistics, 255-288*.
- Ziano, I., & Wang, D. (2021). Slow lies: Response delays promote perceptions of insincerity. *Journal of Personality and Social Psychology, 120*(6), 1457.