Development of language in the minds and brains of children

by
Halie A. Olson

A.B.
Harvard College, 2017

Submitted to the Department of Brain and Cognitive Sciences in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN NEUROSCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2023

Authored by: Halie A. Olson
Department of Brain and Cognitive Sciences
April 3, 2023

Certified by: Rebecca R. Saxe
John W Jarve (1978) Professor of Cognitive Neuroscience
Thesis Supervisor

John D. E. Gabrieli
Grover Hermann Professor of Health Sciences and Technology and Cognitive Neuroscience
Thesis Supervisor

Accepted by: Mark Harnett
Graduate Officer, Department of Brain and Cognitive Sciences

Development of language in the minds and brains of children

by
Halie A. Olson

Submitted to the Department of Brain and Cognitive Sciences on April 3, 2023 in Partial Fulfillment of the Requirements for the Degree of DOCTOR OF PHILOSOPHY IN NEUROSCIENCE

ABSTRACT

One of the most remarkable aspects of early childhood is language development: a process that epitomizes the interplay between nature and nurture in the human brain. Language skills blossom during the second year of life, but they continue to be shaped by children's experiences as they grow and encounter new words and linguistic structures through formal education and their own pursuits. In this thesis, I present three lines of research that aimed to characterize and explore various influences on language processing in children, using both neural and behavioral measures. The first set of studies (Chapters 2-3) involved the development and validation of a functional magnetic resonance imaging (fMRI) task designed to measure language-evoked activation in the brains of awake toddlers, an under-studied age group in fMRI research that represents a critical period of time in language development. Using this task in adults, we found no difference in canonical language regions' responses to speech in dialogue compared to monologue. Ongoing work in toddlers suggests that we can measure language-evoked activation with our approach, which will enable us to characterize language network function for the first time in awake toddlers using fMRI, as well as better understand how social context may or may not impact language processing. The second study (Chapter 4) investigated how children's personal interests impact language processing in the brain. In both neurotypical and autistic children with strong interests, activation in canonical language regions was significantly greater when they listened to personalized stories about their interests than when they listened to generic stories, pointing to the importance of content on the brain's response to language in childhood. Finally, the last study (Chapters 5-6) implemented a remote, randomized controlled trial intervention to examine the impact of audiobooks paired with instructional support on children's language skills. Using vocabulary measures tailored to the books children read, results suggest that struggling readers improved only when audiobooks were paired with instructional support. Together, these studies introduce novel approaches to measuring language in

the minds and brains of children, and explore how factors such as interest and exposure may impact language processing during development.

Thesis Supervisor:	Rebecca R. Saxe
Title:	John W Jarve (1978) Professor of Cognitive Neuroscience
Thesis Supervisor:	John D. E. Gabrieli
Title:	Grover Hermann Professor of Health Sciences and Technology and Cognitive Neuroscience

Acknowledgements

I am so grateful for the many people who got me here today.

First, to my staunch advisors, who let me feel like I was still forging my own path while carefully guiding me along the way: John Gabrieli, thank you for your unending calm positivity and support, and Rebecca Saxe, thank you for both literally and figuratively kayaking beside me through the waves with such enthusiasm and joy.

Second, to my committee members, for providing feedback at critical junctures of my academic journey: Laura Schulz and Marina Bedny.

Third, to my co-first authors on some of the following work, for collaborating with me and making science better and more enjoyable than doing it alone: Anila D'Mello and Kristy Johnson on the Project on Affinities and Language (PAL; Chapter 4) and Ola Ozernov-Palchik on the Audiobook Learning Initiative (ALI; Chapters 5-6).

Fourth, to my other incredible project teammates, for making this research happen. The Language in Toddlers (LIT) squad: Emily Chen, Kirsten Lydic, Somaia Saba, Hana Ro, Rebecca O'Connor, Sofia Riskin, and Michelle Hung. The PAL team: Isabelle Frosch, Shruti Nishith, Hannah Grotzinger, Cindy Li, Jimmy Chen, Nicole Dundas, Insha Merchant, Rucha Kelkar, Alana Kalehua, and Hillary Jean-Gilles. And the ALI crew: Xochitl Arechiga, Kim Wang, Yesi Camacho Torres, Hope Kentala, Natalie Gardino, Jeff Dieffenbach, Isaac Treves, Cindy Li, Jovita Solorio-Fielder, Tolu Asade, Cherry Wang, Sophia Angus, Bhuvna Murthy, Maycee McClure, Sehyr Khan, Hilary Zen, Sarah Abodalo, Elizabeth Carbonell, Shruti Das, Erika Leasher, Yoon Lim, Emmi Mills, Zoë Elizee, Camille Uldry, Shelby Laitipaya, Alexis Cho, Gabriella Aponte, Harley Yoder, Dana Osei, Niki Kim, Joy Bhattacharya, Amanda Miller, David Bates, Ross Weissman, Joohee Baik, June Okada, William Oliver, Harriet Richards, Kristen Wehara, Brooke Goldstein, Ada Huang, Emily Lin, Avni Ayer, Avery Dolins, and Abigail Cassidy. Over 40 undergraduates and high schoolers worked with me on my research projects (a few on other projects that didn't make it into this dissertation, including Aiyedun Uzamere and Ashti Shah) – thank you, also, for helping me learn how to be a better mentor.

Fifth, to my other labmates in Saxelab and Gablab, for supporting me, teaching me, and making science fun. I have worked on other projects with many of you that did not make it into this thesis, but were just as meaningful and important, including: Hilary

Richardson, Heather Kosakowski, Liron Rozencrantz, Rachel Romeo, and Steven Meisler. A special thank you to those of you who provided feedback on parts of this dissertation, including Sabrina Piccolo. And to everyone who accompanied me to get a chai latte at some point over the last few years, particularly Anila D'Mello, Sadie Zacharek, Emily Chen, and Amanda O'Brien, for fueling my productivity.

Sixth, to many others in the Building 46 community, for making BCS a home during my time here, and for making my science possible: Steven Shannon, Atsushi Takahashi, Jamie Wiley, Laura Frawley, Sierra Vallin, Julianne Ormerod, Kris Brewer, McGovern and BCS HQs, my cohort and other BCS grad students, BCS faculty who asked me hard questions and provided me with thoughtful suggestions - particularly Ev Fedorenko and Nancy Kanwisher - and everyone I interacted with over the past years.

Seventh, to the families who participated in my research projects, for making this possible and worthwhile.

Eighth, to all of my teachers, coaches, mentors, and professors who taught me, encouraged me, and fostered my love of learning, for giving me the skills I would need to succeed before I even started my PhD. In particular, thank you to Dr. Chuck Nelson, who introduced me to the field of developmental cognitive neuroscience and started me on this path. And to my friends, who helped me find balance and fun.

Ninth, to my family, for being there for me from the very beginning: especially my parents, Alison and Paul Olson, and my brothers, Hunter and Max Olson. And to my dog, Winston, who came into my life at the beginning of my PhD journey, for bringing me comfort, joy, and warmth.

And finally, tenth, to my husband and partner, Lee Tarlin, for experiencing the highs and lows of this whole endeavor, for listening to me talk through every decision ad nauseum, and for being my biggest supporter.

My heart is full of love and gratitude for you all. Thank you.

Table of Contents

List of Figures

List of Tables

"I know my mind in terms of a language more expressive than any I'd previously imagined."

-Ted Chiang, "Understand," ***Stories of Your Life and Others***

Chapter 1 : Introduction

Human infants, toddlers, and children are unparalleled when it comes to learning language. They are much better than machines[1] (Dupoux, 2018) and other animals (Hauser et al., 2002; ten Cate & Okanoya, 2012), and also much better than adult humans at learning a native language (Friedmann & Rusou, 2015; Grimshaw et al., 1998; Newport, 1990). While their experiences with language vary immensely, young children learning language will generally progress through common milestones. They will first begin babbling around 7 months, then produce single words around 12 months, and then combine words around 18 months (Hoff, 2013b; Kuhl, 2004). This explosion of language is particularly apparent in toddlerhood, when the number of words children comprehend and produce increases exponentially (Frank et al., 2021).

Despite these remarkably similar end states and overall trajectories, individual paths to language learning throughout childhood are immensely varied. Pick any two infants, and chances are certain that they will not develop language skills in the exact same way. For instance, some children begin speaking their first words at nine months, and some will begin speaking in double that amount of time but then do so fluently. Other children are delayed and never catch up to their peers. Frank and colleagues put it succinctly: "On average, language emerges quickly – but despite the average pattern,

[1] Probably not for long, though.

toddlers around the world are 'all over the place' in their learning rate. Indeed, one of the most compelling universals in language development is the variation that is observed across children" (Frank et al., 2021).

Of course, language learning does not end in toddlerhood, nor does individual variability in language skills. Language development is contingent upon experience, and one of the most profound drivers of linguistic input for many children is formalized education. In particular, learning to read introduces children to richer vocabulary and more complex syntactic structures than everyday speech (Acheson et al., 2008; Duff et al., 2015), as well as to diverse content areas. As in toddlerhood, there are remarkable similarities in language development during childhood, but also individual variability driven by factors ranging from neurodevelopmental disorders and learning difficulties (Bishop & Snowling, 2004; Catts, 1991; Hudry et al., 2010; Pickles et al., 2014; Snowling, 2001) to socioeconomic environment (Fernald et al., 2013; Hoff, 2013a; Walker et al., 1994).

In this thesis, I present a research program that aims to characterize and explore varied influences on language learning in development, both neurally and behaviorally. The studies described here grapple with profound challenges arising from the fact that language learning is inevitably variable based on both endogenous and exogenous variability, and thus that measuring language competence by measuring language responses in either the brain or behavior is fraught. Developmentally-informed measurements of language need to respond to the fact that language learning is dependent upon interest, experience, content, modality, and a host of other factors. One way to measure language learning is to measure activation of brain regions. But a challenge is: can one robustly and reliably measure brain activation elicited by

language comprehension in very young children? And another challenge is: can we meaningfully isolate language as a construct from the content of the materials?

This thesis will explore two primary themes. The first theme is **innovating techniques for measuring language activation in the brain in difficult-to-reach developmental populations**. This theme is most prevalent in chapters 2-4, which describe functional magnetic resonance imaging (fMRI) paradigms designed to measure brain activation during language comprehension in toddlers and children, including autistic[2] children. The second theme is **conceptualizing what it means to measure language competence while accounting for variability in children's interests and experiences**. This theme is most prevalent in chapters 4-6, which describe the impact of personal interest on language network response in the brain, as well as the impact of listening to audiobooks on different measures of vocabulary learning. In this Introduction, I will outline each of these themes, and articulate the gaps that each study aimed to address.

Measuring language development in the brain

Arguing for the necessity of innovation to study language activation in the brains of difficult-to-reach developmental populations is predicated on an assumption that it is worthwhile – perhaps even necessary – to study the brain to understand this

[2] Some individuals prefer identity-first language (e.g., "autistic individuals") whereas other prefer person-first language (e.g., "individuals with autism") (Amaral, 2023; Robison, 2019). In this thesis, I will primarily use identity-first language as this was endorsed more than person-first language by autistic adults in recent work (Keating et al., 2023).

developmental process. This is certainly debatable (e.g., Francken et al., 2022), but I'll posit three reasons why I find it a compelling approach[3]:

1. Brain measures may be 'closer to the source' of true underlying differences in language competence. Because they rely on other functions (like attention, motivation, compliance, etc.), behavioral measures can mis- and under-estimate cognitive capacities, particularly in young and neurodiverse populations (e.g., Bates, 1993; Bloom & German, 2000; Onishi & Baillargeon, 2005; Saxe et al., 2004).
2. The brain allows us to measure 'something else.' By no means should brain measures replace behavioral measures, but they can complement our understanding by elucidating quantitative functional and structural properties that can be measured and tracked over time and across individuals (e.g., Hoeft et al., 2011).
3. Because we know (a fair amount) about the functional underpinnings of language in the adult brain, we can compare measured properties in developing brains to mature ones (e.g., Szaflarski et al., 2006).

With these justifications in mind, I will now move on how we study language in the brain, and why we need to adapt this approach when considering certain developmental populations.

[3] Other people also think it is worth studying the developing brain for lots of reasons, such as (Dehaene-Lambertz, 2017).

Neural basis of language in the adult brain

The most common approach to measuring the brain's response to 'language' using fMRI is to subtract activation evoked by a control condition from a condition of interest, usually in a block design, such as: auditory speech versus acoustically-degraded speech (e.g., Overath et al., 2015; Scott et al., 2017; Stoppelman et al., 2013), foreign speech (e.g., Schlosser et al., 1998), or backwards speech (e.g., Bedny et al., 2011; Dehaene-Lambertz et al., 2002; Moore-Parks et al., 2010; Redcay et al., 2008); or printed sentences versus lists of nonwords (e.g., Fedorenko et al., 2010). Critically, the control condition is designed to account for all features of the stimulus aside from the key cognitive construct interest. Using such approaches, it has been shown that the adult brain processes language similarly across individuals, exhibiting consistent patterns of response to language stimuli in predominantly left frontal and temporal cortical, and right cerebellar, regions (e.g., Fedorenko et al., 2010, 2011; Friederici, 2011; Friederici & Gierhan, 2013; Price, 2010). These regions respond no matter what language one speaks (Malik-Moraleda et al., 2022), and no matter what modality that language is presented in (e.g., Bedny et al., 2011; Fedorenko et al., 2010; MacSweeney et al., 2008; Neville et al., 1998; Scott et al., 2017). These regions are necessary for language processing once language networks have developed, insofar as damage to these regions often cause language deficits[4] (Broca, 1865; Wernicke, 1874). While there is general agreement that these regions are fundamentally involved in language processing, defining their scope, limits, and specificity has generated more debate[5]

[4] Another fascinating question is how language network develops when these regions are unavailable, due to early damage or congenital brain abnormalities (e.g., (François et al., 2021; Tuckute et al., 2022)).
[5] This debate is not restricted to the domain of language; see (Kanwisher, 2010).

(e.g., Fedorenko & Kanwisher, 2009; Fedorenko & Thompson-Schill, 2014; Grodzinsky, 2010; Monti et al., 2012).

A challenge to probing the functional profile of language network has been figuring out how to define the boundaries of these regions. This has not only been relevant to studying language, but also for studying other cognitive functions in the human brain (Kanwisher, 2010; Saxe et al., 2006). Just like any two faces have roughly similar features – usually, two eyes, a nose, a mouth, etc. – brains have remarkably similar structural and functional architectures. Yet the precise location and shape of these features differ[6]. When brains are averaged together, the boundaries can blur, leading to uncertainty about whether a given area subserves multiple overlapping functions, or whether those functions are close together but spatially distinct. A solution to this problem is to map out functions within individuals – that is, functionally localize a particular region of interest (Saxe et al., 2006) – and then use these individual regions in independent analyses. Functional localizers have been developed for multiple cognitive functions, such as face processing (Kanwisher et al., 1997), theory of mind (Saxe & Kanwisher, 2003), and language (Fedorenko et al., 2010).

Using functionally-localized language regions, it has been shown that language regions are not engaged in non-linguistic cognitive tasks, such as math, working memory, music, or computer coding (Fedorenko et al., 2011; Ivanova et al., 2020; Liu et al., 2020). Language regions are, however, involved in both production and comprehension of language (Hagoort, 2014; Hu et al., 2022; Menenti et al., 2011;

[6] This analogy comes from Ev Fedorenko. I found it incredibly intuitive, so I have continued to use it since.

Price, 2010), and activation in these regions can be modulated by linguistic features, such as complexity (Blank et al., 2016; Hagoort & Indefrey, 2014; Wehbe et al., 2021). Though language is often used in a social context and for social purposes, language network is distinct from theory of mind, even though these networks can sometimes be correlated (Paunov et al., 2019, 2022; Shain et al., 2022).

Why, one might be asking at this point, should we care about the precise scope and limits of these regions? Is it just splitting hairs for the sake of scientific argument?

Others can offer many reasons, such as speaking to debates about the uniqueness of the cognitive processing involved in language comprehension (e.g., Blank et al., 2014; Fedorenko et al., 2011; January et al., 2009). In the context of this work, understanding what language regions do – and do not – represent can constrain our hypotheses about the development of language network. For instance, there is a fundamental interplay between language and social function in everyday life, but there is functional separation between language and social networks in the adult brain. And yet, social context is necessary for supporting language learning in development (Kuhl, 2007). Is this reflected in the functional profile of developing language network? Does language network become specialized over time? Are non-language regions initially involved in language processing, or are additional regions specialized for language at particular timepoints in development? Defining the scopes and limits of the adult language network is critical for interpreting developmental findings.

Neural basis of language in the developing brain

Studies of children in many ways recapitulate the adult work: typically, the same left-lateralized set of cortical regions responds to language stimuli in children (Berl et al., 2010a, 2014; Enge et al., 2020; Gaillard et al., 2001; Holland et al., 2001, 2007; Lidzba et al., 2011; Szaflarski et al., 2006, 2012; Wood et al., 2004). Even in relatively young children, this network appears to be specialized and distinct from multiple demand regions (Hiersche et al., 2022). There is debate, however, regarding how this network changes over time (Olulade et al., 2020).

Most fMRI work[7] has focused on children over 3 years of age due to methodological constraints. However, the most profound changes in language development (Frank et al., 2021), and the period of time in which language development is most sensitive to environmental input (Friedmann & Rusou, 2015), takes place before 3 years of age. There are some exceptions: in particular, studies which have examined *sleeping* infants' and toddlers' responses to auditory speech and other sounds (Dehaene-Lambertz et al., 2002; Kosakowski et al., 2023; Perani et al., 2010, 2011; Redcay et al., 2008; Redcay & Courchesne, 2008; Wild et al., 2017). Sleeping infants and toddlers generally show bilateral and left-lateralized activation in response to speech. However, these responses are not measuring language comprehension, but rather speech perception, and therefore provide limited insight into the developmental origins of language processing in the brain. Intriguingly, in a few babies that woke up in the scanner in Dehaene-Lambertz et al's seminal study, there was frontal activation that was not seen

[7] There has been a large body of functional neuroimaging work on language processing and other cognitive functions in this age group using other methodologies, such as functional near-infrared spectroscopy (fNIRS) and electroencephalography (EEG). These methods are superior to fMRI for young children in many ways, but do not have the spatial resolution to localize cognitive functions in the brain.

in the sleeping infants (Dehaene-Lambertz et al., 2002), and in a second study in awake infants, some modulation of frontal activation to repeated sentences was observed (Dehaene-Lambertz et al., 2006). The specific effects of sleep on functional activation – particularly in young children – are not well understood (Cusack et al., 2018; Mitra et al., 2017); thus, awake task-based fMRI is an important tool for understanding the developing brain before age 3 years (Deen et al., 2017; Ellis et al., 2020; Ellis & Turk-Browne, 2018; Kosakowski et al., 2022; Yates et al., 2021a, 2021b).

Developmental approaches to studying language in the brain

In order to most effectively estimate brain function, we need experimental paradigms that will drive the neural activity of – and also hold the attention of – the population we are aiming to study. To study language function in adult brains, we can ask participants to read sentence after sentence flashing up on a white screen, and (1) they will (usually) do what we ask, and (2) this paradigm will drive a reliable and robust response in predictable patterns in their brains. This approach will not work for a two-year-old.

While the above example was intentionally set up as a strawman, many of the approaches we use to study brain function in adults simply do not work for young children or neurodiverse populations. Young children have shorter attention spans, move more, lack some of the cognitive skills that adults have (like the ability to read), and may be less compliant than older participants, which introduce challenges for functional neuroimaging (Poldrack et al., 2002; Raschle et al., 2012). These are not just annoying challenges for keeping a child in the scanner – they can be critical for determining which inferences we can make about data collected from these populations. For instance, do autistic children show less activation in a particular task

than neurotypical children because these regions are less engaged, or because they were less interested in the stimuli? Is resting state functional connectivity lower because the networks are still developing, or because younger children move more (e.g., Satterthwaite et al., 2012)?

For developmental studies of language, the ideal stimuli should:

1. Be engaging for the target age range, across the entire functional run, across participants.
2. Drive neural activity for the effect of interest in a wide age range, if looking for developmental change.
3. Be sensitive to differences in other cognitive skills that are simultaneously emerging.

Researchers have developed a number of approaches to address some of these challenges. For instance, using naturalistic stimuli – like movies or stories – is a promising approach for studying brain function, especially in children (Cantlon, 2020; Cantlon & Li, 2013; Kamps et al., 2022; Redcay & Moraczewski, 2020; Richardson et al., 2018; Vanderwal et al., 2015, 2019). Playing videos can decrease child motion during scans while robustly eliciting neural responses (Frew et al., 2021). A drawback, however, is that certain cognitive processes may not be isolated in commercially produced movies; when studying language, for example, most movies do not have a non-speech control built in (see our approach to introducing this control in Chapters 2-3).

Indeed, we often focus on controlling lower-level aspects of stimuli, like acoustic properties, at the expense of other higher-level potential confounds. However, we

know that these other properties matter. When measuring language comprehension, for example, children perform better when the materials are personally interesting and familiar to them (Baldwin et al., 1985; Shnayer, 1968). Studies of reward network have used personalized stimuli to maximize relevance to the individual (e.g., Kohls et al., 2018; Tomova et al., 2020), and even studies of face processing have shown effects of familiarity (Pierce & Redcay, 2008). Language is shaped by experience, and we thus cannot study language as if it were unrelated to the interests of an individual (see Chapter 4) or their exposure to certain materials (see Chapter 5-6). A one-size-fits-all approach with generic measures may underestimate the capacity of brain networks in certain individuals, which can be particularly problematic for studies of development, individual or group differences, or intervention response.

Endogenous and exogenous influences on language measurements

This brings us to the second theme: when we are trying to measure language competence, how do we distinguish between language-specific responses and the attention, familiarity, interest, background knowledge, and other factors that impact how particular language stimuli are processed?

Individual differences in language processing

There are true individual differences in language processing, reflected in both behavioral measures and brain activation. Sometimes, these differences are driven by particular experiences or biological factors that can be identified. For example, occipital cortex contributes to language processing in congenitally blind individuals (Bedny et al., 2011), a difference from language processing in sighted individuals that

is driven by a specific environmental input during the development of this network. Yet in other cases, the factors that drive differences in language as measured by the brain and behavior are less clear.

For example, endogenous variability in the form of neurodevelopmental and learning disorders, such as autism spectrum disorder or dyslexia, respectively, are associated with variation in both a child's language skills *and* their linguistic experiences (Bishop & Snowling, 2004; Catts, 1991; Hudry et al., 2010; Pickles et al., 2014; Snowling, 2001). Thus, while variability in language-evoked brain activation can be correlated with diagnoses (e.g., autism, dyslexia; Herringshaw et al., 2016; Mody & Belliveau, 2013; Shaywitz et al., 1998), it is difficult to identify the mechanism or cause of these differences. Variation in brain activation to language also exists within non-clinical populations, and may relate to language skills (e.g., Berl et al., 2010b; Sroka et al., 2015) – but again, the impact of endogenous and exogenous factors is difficult to untangle. For example, some evidence suggests that certain language experiences, such as reading books with complex language (Acheson et al., 2008; Duff et al., 2015) and engaging in child-directed speech and back-and-forth conversations (Ferjan Ramírez et al., 2020; Hirsh-Pasek et al., 2015; Romeo et al., 2018; Rowe, 2008), may positively impact language skills and trajectories. Indeed, the communicative and social aspects of language are vitally important to language learning and development (Hoff, 2006; Kuhl, 2004; Kuhl et al., 2003). Identifying the impact of different factors, in the brain and behavior, is critical for understanding language development – but in order to do so, it is also critical to have robust and reliable measures of language itself.

Addressing individual differences in language measurements

Language cannot be separated from either content or context – comprehending language involves relating the visual and auditory signals to meaning, and that meaning is informed by experience. That also means that our measures of "language" invariably pick up on other factors. Children's reading comprehension, for example, is better when they read something they are familiar with or interested in (Baldwin et al., 1985; Shnayer, 1968), and measures of vocabulary have long been biased due to the populations they are normed on and the items included (Kachergis et al., 2022; Restrepo et al., 2006). It can be challenging to get reliable estimates of language abilities from young children in research settings for many reasons (Dockrell & Marshall, 2015); for instance, young children may not comply with instructions, may be shy, or may not pay attention. Parent reports can help (Dale et al., 1989; Law & Roy, 2008), but they are also influenced by characteristics of the parents filling them out (Feldman et al., 2000; Pan et al., 2004). One might assume that brain measures are less biased, but the same factors may be affecting measures of brain function – how interesting, engaging, or familiar the language is.

Rather than accepting these confounding factors, one approach is to lean into them through personalization in study design. If children find different topics particularly interesting, then use stimuli based on those interests to measure more robust effects (see Chapter 4). If children are more motivated to read certain books than others, provide that choice, and modify the assessments accordingly (see Chapter 5-6). This approach is of course used in classrooms all the time, but can be more difficult to justify from a research standpoint when the goal is to standardize measurements as

much as possible. The tradeoff, however, may be more reliable measurements of the underlying capacity we wish to study.

An overview of the following chapters

This dissertation includes four studies that use diverse experimental paradigms to better understand the neural basis of language processing.

The first three chapters focus on **innovative techniques** for measuring language-evoked activation in the brain. **Chapter 2** introduces an fMRI paradigm that embeds experimental control within engaging, child-friendly naturalistic videos. Using this task in adults, we showed that canonical left-hemisphere cortical language regions do not differentiate between dialogue and speech coming from a single source. This study contributes to a large body of literature on the scope and limits of language network function, confirming the dissociation between local language processing and the context in which language is encountered in the adult brain. In addition to probing the scope of language network function, this study also served as a validation of a novel set of language localizer tasks designed for young children, which were used in the study described in Chapter 3.

To understand the neural basis of language development, a critical step is to measure the brain's functional response to language during a period of rapid and remarkable changes in language development: toddlerhood. **Chapter 3** describes how we used the tasks described in Chapter 2 to scan awake toddlers, an age group that is notoriously difficult to scan using fMRI. While preliminary, these results suggest we can identify language-evoked activation in left-lateralized language regions in toddlers.

The study described in Chapter 4 once again probes the sensitivity of language network, in this case to the content of language, through personalized study design. Children – and adults – often have highly focused interests in specific topics, which can motivate how they spend their free time, who they form relationships with, and what they talk about. In the scanner, children listened to personalized language stimuli written about their specific interest. Language network responded more to language about their interest than non-personalized stories about nature. Autistic individuals often have highly restricted special interests that can interfere with daily life, but can also motivate social interaction and communication (Baker et al., 1998; Cascio et al., 2014; Charlop-Christy & Haymes, 1998; Klin et al., 2007). Like the neurotypical children, autistic children had higher responses to stories about their interest than non-personalized stories in language network.

Like Chapter 4, the final two chapters call into question whether standardization may sometimes limit our ability to measure the full extent of language skills in certain populations. Specifically, Chapters 5 and 6 detail a remote randomized controlled trial (RCT) investigating the effects of an audiobook intervention on children's language skills, conducted during the COVID-19 pandemic. When tested on standardized measures of vocabulary, children in the intervention groups did not seem to improve relative to the control group, but when tested using measures tailored to the specific books each child listened to, poor readers in one of the intervention groups did show vocabulary gains in preliminary exploratory analyses.

In sum, the following chapters explore the minds, brains, interests, expressions, and language of over 300 toddlers, children, and adults who spent their time – much of it

during a pandemic – contributing to the scientific endeavor to understand the human brain and its incredible capacity for language. In Chapter 7, I will discuss the implications of these studies together, and how they may inform future studies of language processing and development in children.

References

Acheson, D. J., Wells, J. B., & MacDonald, M. C. (2008). New and updated tests of print exposure and reading abilities in college students. *Behavior Research Methods*, *40*(1), 278–289. https://doi.org/10.3758/BRM.40.1.278

Amaral, D. G. (2023). Language in Autism Research: Accurate and Respectful. *Autism Research*, *16*(1), 7–8. https://doi.org/10.1002/aur.2886

Baker, M. J., Koegel, R. L., & Koegel, L. K. (1998). Increasing the Social Behavior of Young Children with Autism Using Their Obsessive Behaviors. *Journal of the Association for Persons with Severe Handicaps*, *23*(4), 300–308. https://doi.org/10.2511/rpsd.23.4.300

Baldwin, R. S., Peleg-Bruckner, Z., & McClintock, A. H. (1985). Effects of Topic Interest and Prior Knowledge on Reading Comprehension. *Reading Research Quarterly*, *20*(4), 497–504. https://doi.org/10.2307/747856

Bates, E. (1993). Comprehension and Production in Early Language Development. *Monographs of the Society for Research in Child Development*, *58*(3–4), 222–242. https://doi.org/10.1111/j.1540-5834.1993.tb00403.x

Bedny, M., Pascual-Leone, A., Dodell-Feder, D., Fedorenko, E., & Saxe, R. (2011). Language processing in the occipital cortex of congenitally blind adults. *Proceedings of the National Academy of Sciences*, *108*(11), 4429–4434. https://doi.org/10.1073/pnas.1014818108

Berl, M. M., Duke, E. S., Mayo, J., Rosenberger, L. R., Moore, E. N., VanMeter, J., Ratner, N. B., Vaidya, C. J., & Gaillard, W. D. (2010a). Functional anatomy of listening and reading comprehension during development. *Brain and Language*, *114*(2), 115–125. https://doi.org/10.1016/j.bandl.2010.06.002

Berl, M. M., Duke, E. S., Mayo, J., Rosenberger, L. R., Moore, E. N., VanMeter, J., Ratner, N. B., Vaidya, C. J., & Gaillard, W. D. (2010b). Functional anatomy of listening and reading comprehension during development. *Brain and Language*, *114*(2), 115–125. https://doi.org/10.1016/j.bandl.2010.06.002

Berl, M. M., Mayo, J., Parks, E. N., Rosenberger, L. R., VanMeter, J., Ratner, N. B., Vaidya, C. J., & Gaillard, W. D. (2014). Regional differences in the

developmental trajectory of lateralization of the language network. *Human Brain Mapping*, *35*(1), 270–284. https://doi.org/10.1002/hbm.22179

Bishop, D. V. M., & Snowling, M. J. (2004). Developmental Dyslexia and Specific Language Impairment: Same or Different? *Psychological Bulletin*, *130*, 858–886. https://doi.org/10.1037/0033-2909.130.6.858

Blank, I., Balewski, Z., Mahowald, K., & Fedorenko, E. (2016). Syntactic processing is distributed across the language system. *NeuroImage*, *127*, 307–323. https://doi.org/10.1016/j.neuroimage.2015.11.069

Blank, I., Kanwisher, N., & Fedorenko, E. (2014). A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. *Journal of Neurophysiology*, *112*(5), 1105–1118. https://doi.org/10.1152/jn.00884.2013

Bloom, P., & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, *77*(1), B25–B31. https://doi.org/10.1016/S0010-0277(00)00096-2

Broca, P. (1865). Sur le siège de la faculté du langage articulé. *Bulletins et Mémoires de la Société d'Anthropologie de Paris*, *6*(1), 377–393. https://doi.org/10.3406/bmsap.1865.9495

Cantlon, J. F. (2020). The balance of rigor and reality in developmental neuroscience. *NeuroImage*, *216*, 116464. https://doi.org/10.1016/j.neuroimage.2019.116464

Cantlon, J. F., & Li, R. (2013). Neural Activity during Natural Viewing of Sesame Street Statistically Predicts Test Scores in Early Childhood. *PLOS Biology*, *11*(1), e1001462. https://doi.org/10.1371/journal.pbio.1001462

Cascio, C. J., Foss-Feig, J. H., Heacock, J., Schauder, K. B., Loring, W. A., Rogers, B. P., Pryweller, J. R., Newsom, C. R., Cockhren, J., Cao, A., & Bolton, S. (2014). Affective neural response to restricted interests in autism spectrum disorders. *Journal of Child Psychology and Psychiatry*, *55*(2), 162–171. https://doi.org/10.1111/jcpp.12147

Catts, H. W. (1991). Early Identification of Dyslexia: Evidence from a Follow-Up Study of Speech-Language Impaired Children. *Annals of Dyslexia*, *41*, 163–177.

Charlop-Christy, M. H., & Haymes, L. K. (1998). Using Objects of Obsession as Token Reinforcers for Children with Autism. *Journal of Autism and Developmental Disorders*, *28*(3), 189–198. https://doi.org/10.1023/A:1026061220171

Cusack, R., McCuaig, O., & Linke, A. C. (2018). Methodological challenges in the comparison of infant fMRI across age groups. *Developmental Cognitive Neuroscience*, *33*, 194–205. https://doi.org/10.1016/j.dcn.2017.11.003

Dale, P. S., Bates, E., Reznick, J. S., & Morisset, C. (1989). The validity of a parent report instrument of child language at twenty months. *Journal of Child Language*, *16*(2), 239–249. https://doi.org/10.1017/S0305000900010394

Deen, B., Richardson, H., Dilks, D. D., Takahashi, A., Keil, B., Wald, L. L., Kanwisher, N., & Saxe, R. (2017). Organization of high-level visual cortex in human infants. *Nature Communications*, *8*(1), Article 1. https://doi.org/10.1038/ncomms13995

Dehaene-Lambertz, G. (2017). The human infant brain: A neural architecture able to learn language. *Psychonomic Bulletin & Review*, *24*(1), 48–55. https://doi.org/10.3758/s13423-016-1156-9

Dehaene-Lambertz, G., Dehaene, S., & Hertz-Pannier, L. (2002). Functional Neuroimaging of Speech Perception in Infants. *Science*, *298*(5600), 2013–2015. https://doi.org/10.1126/science.1077066

Dehaene-Lambertz, G., Hertz-Pannier, L., Dubois, J., Mériaux, S., Roche, A., Sigman, M., & Dehaene, S. (2006). Functional organization of perisylvian activation during presentation of sentences in preverbal infants. *Proceedings of the National Academy of Sciences*, *103*(38), 14240–14245. https://doi.org/10.1073/pnas.0606302103

Dockrell, J. E., & Marshall, C. R. (2015). Measurement Issues: Assessing language skills in young children. *Child and Adolescent Mental Health*, *20*(2), 116–125. https://doi.org/10.1111/camh.12072

Duff, D., Tomblin, J. B., & Catts, H. (2015). The Influence of Reading on Vocabulary Growth: A Case for a Matthew Effect. *Journal of Speech, Language, and Hearing Research*, *58*(3), 853–864. https://doi.org/10.1044/2015_JSLHR-L-13-0310

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, *173*, 43–59. https://doi.org/10.1016/j.cognition.2017.11.008

Ellis, C. T., Skalaban, L. J., Yates, T. S., Bejjanki, V. R., Córdova, N. I., & Turk-Browne, N. B. (2020). Re-imagining fMRI for awake behaving infants. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-18286-y

Ellis, C. T., & Turk-Browne, N. B. (2018). Infant fMRI: A Model System for Cognitive Neuroscience. *Trends in Cognitive Sciences*, *22*(5), 375–387. https://doi.org/10.1016/j.tics.2018.01.005

Enge, A., Friederici, A. D., & Skeide, M. A. (2020). A meta-analysis of fMRI studies of language comprehension in children. *NeuroImage*, *215*, 116858. https://doi.org/10.1016/j.neuroimage.2020.116858

Fedorenko, E., Behr, M. K., & Kanwisher, N. (2011). Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences*, *108*(39), 16428–16433. https://doi.org/10.1073/pnas.1112937108

Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New Method for fMRI Investigations of Language: Defining ROIs Functionally in Individual Subjects. *Journal of Neurophysiology*, *104*(2), 1177–1194. https://doi.org/10.1152/jn.00032.2010

Fedorenko, E., & Kanwisher, N. (2009). Neuroimaging of Language: Why Hasn't a Clearer Picture Emerged? *Language and Linguistics Compass*, *3*(4), 839–865. https://doi.org/10.1111/j.1749-818X.2009.00143.x

Fedorenko, E., & Thompson-Schill, S. L. (2014). Reworking the language network. *Trends in Cognitive Sciences*, *18*(3), 120–126. https://doi.org/10.1016/j.tics.2013.12.006

Feldman, H. M., Dollaghan, C. A., Campbell, T. F., Kurs-Lasky, M., Janosky, J. E., & Paradise, J. L. (2000). Measurement Properties of the MacArthur Communicative Development Inventories at Ages One and Two Years. *Child Development*, *71*(2), 310–322. https://doi.org/10.1111/1467-8624.00146

Ferjan Ramírez, N., Lytle, S. R., & Kuhl, P. K. (2020). Parent coaching increases conversational turns and advances infant language development. *Proceedings of the National Academy of Sciences*, *117*(7), 3484–3491. https://doi.org/10.1073/pnas.1921653117

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science*, *16*(2), 234–248. https://doi.org/10.1111/desc.12019

Francken, J. C., Slors, M., & Craver, C. F. (2022). Cognitive ontology and the search for neural mechanisms: Three foundational problems. *Synthese*, *200*(5), 378. https://doi.org/10.1007/s11229-022-03701-2

François, C., Garcia-Alix, A., Bosch, L., & Rodriguez-Fornells, A. (2021). Signatures of brain plasticity supporting language recovery after perinatal arterial ischemic stroke. *Brain and Language*, *212*, 104880. https://doi.org/10.1016/j.bandl.2020.104880

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and Consistency in Early Language Learning: The Wordbank Project*. MIT Press.

Frew, S., Samara, A., Shearer, H., Eilbott, J., & Vanderwal, T. (2021). *Getting the Nod: Characterizing pediatric head motion in movie- and resting-state fMRI* [Preprint]. Radiology and Imaging. https://doi.org/10.1101/2021.05.17.21257346

Friederici, A. D. (2011). The Brain Basis of Language Processing: From Structure to Function. *Physiological Reviews*, *91*(4), 1357–1392. https://doi.org/10.1152/physrev.00006.2011

Friederici, A. D., & Gierhan, S. M. (2013). The language network. *Current Opinion in Neurobiology*, *23*(2), 250–254. https://doi.org/10.1016/j.conb.2012.10.002

Friedmann, N., & Rusou, D. (2015). Critical period for first language: The crucial role of language input during the first year of life. *Current Opinion in Neurobiology*, *35*, 27–34. https://doi.org/10.1016/j.conb.2015.06.003

Gaillard, W. D., Pugliese, M., Grandin, C. B., Braniecki, S. H., Kondapaneni, P., Hunter, K., Xu, B., Petrella, J. R., Balsamo, L., & Basso, G. (2001). Cortical localization of reading in normal children: An fMRI language study. *Neurology*, *57*(1), 47–54. https://doi.org/10.1212/WNL.57.1.47

Grimshaw, G. M., Adelstein, A., Bryden, M. P., & MacKinnon, G. E. (1998). First-Language Acquisition in Adolescence: Evidence for a Critical Period for Verbal Language Development. *Brain and Language*, *63*(2), 237–255. https://doi.org/10.1006/brln.1997.1943

Grodzinsky, Y. (2010). The Picture of the Linguistic Brain: How Sharp Can It Be? Reply to Fedorenko & Kanwisher. *Language and Linguistics Compass*, *4*(8), 605–622. https://doi.org/10.1111/j.1749-818X.2010.00222.x

Hagoort, P. (2014). Nodes and networks in the neural architecture for language: Broca's region and beyond. *Current Opinion in Neurobiology*, *28*, 136–141. https://doi.org/10.1016/j.conb.2014.07.013

Hagoort, P., & Indefrey, P. (2014). The Neurobiology of Language Beyond Single Words. *Annual Review of Neuroscience*, *37*(1), 347–362. https://doi.org/10.1146/annurev-neuro-071013-013847

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). *The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? 298*.

Herringshaw, A. J., Ammons, C. J., DeRamus, T. P., & Kana, R. K. (2016). Hemispheric differences in language processing in autism spectrum disorders: A meta-analysis of neuroimaging studies. *Autism Research*, *9*(10), 1046–1057. https://doi.org/10.1002/aur.1599

Hiersche, K. J., Schettini, E., Li, J., & Saygin, Z. M. (2022). *The language network is selective and distinct from other cognition in both function and connectivity in early childhood* (p. 2022.08.11.503597). bioRxiv. https://doi.org/10.1101/2022.08.11.503597

Hirsh-Pasek, K., Adamson, L. B., Bakeman, R., Owen, M. T., Golinkoff, R. M., Pace, A., Yust, P. K. S., & Suma, K. (2015). The Contribution of Early Communication Quality to Low-Income Children's Language Success. *Psychological Science*, *26*(7), 1071–1083. https://doi.org/10.1177/0956797615581493

Hoeft, F., McCandliss, B. D., Black, J. M., Gantman, A., Zakerani, N., Hulme, C., Lyytinen, H., Whitfield-Gabrieli, S., Glover, G. H., Reiss, A. L., & Gabrieli, J. D. E. (2011). Neural systems predicting long-term outcome in dyslexia. *Proceedings*

of the National Academy of Sciences, *108*(1), 361–366. https://doi.org/10.1073/pnas.1008950108

Hoff, E. (2006). How social contexts support and shape language development. *Developmental Review*, *26*(1), 55–88. https://doi.org/10.1016/j.dr.2005.11.002

Hoff, E. (2013a). Interpreting the early language trajectories of children from low-SES and language minority homes: Implications for closing achievement gaps. *Developmental Psychology*, *49*, 4–14. https://doi.org/10.1037/a0027238

Hoff, E. (2013b). *Language Development*. Cengage Learning.

Holland, S. K., Plante, E., Weber Byars, A., Strawsburg, R. H., Schmithorst, V. J., & Ball, W. S. (2001). Normal fMRI Brain Activation Patterns in Children Performing a Verb Generation Task. *NeuroImage*, *14*(4), 837–843. https://doi.org/10.1006/nimg.2001.0875

Holland, S. K., Vannest, J., Mecoli, M., Jacola, L. M., Tillema, J.-M., Karunanayaka, P. R., Schmithorst, V. J., Yuan, W., Plante, E., & Byars, A. W. (2007). Functional MRI of language lateralization during development in children. *International Journal of Audiology*, *46*(9), 533–551. https://doi.org/10.1080/14992020701448994

Hu, J., Small, H., Kean, H., Takahashi, A., Zekelman, L., Kleinman, D., Ryan, E., Nieto-Castañón, A., Ferreira, V., & Fedorenko, E. (2022). Precision fMRI reveals that the language-selective network supports both phrase-structure building and lexical access during language production. *Cerebral Cortex*, bhac350. https://doi.org/10.1093/cercor/bhac350

Hudry, K., Leadbitter, K., Temple, K., Slonims, V., McConachie, H., Aldred, C., Howlin, P., & Charman, T. (2010). Preschoolers with autism show greater impairment in receptive compared with expressive language abilities. *International Journal of Language & Communication Disorders*, *45*(6), 681–690. https://doi.org/10.3109/13682820903461493

Ivanova, A. A., Srikant, S., Sueoka, Y., Kean, H. H., Dhamala, R., O'Reilly, U.-M., Bers, M. U., & Fedorenko, E. (2020). Comprehension of computer code relies primarily on domain-general executive brain regions. *ELife*, *9*, e58906. https://doi.org/10.7554/eLife.58906

January, D., Trueswell, J. C., & Thompson-Schill, S. L. (2009). Co-localization of Stroop and Syntactic Ambiguity Resolution in Broca's Area: Implications for the Neural Basis of Sentence Processing. *Journal of Cognitive Neuroscience*, *21*(12), 2434–2444. https://doi.org/10.1162/jocn.2008.21179

Kachergis, G., Francis, N., & Frank, M. C. (2022). Estimating Demographic Bias on Tests of Children's Early Vocabulary. *Topics in Cognitive Science*, *n/a*(n/a). https://doi.org/10.1111/tops.12635

Kamps, F. S., Richardson, H., Murty, N. A. R., Kanwisher, N., & Saxe, R. (2022). Using child-friendly movie stimuli to study the development of face, place, and object regions from age 3 to 12 years. *Human Brain Mapping*, *43*(9), 2782–2800. https://doi.org/10.1002/hbm.25815

Kanwisher, N. (2010). Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences*, *107*(25), 11163–11170. https://doi.org/10.1073/pnas.1005062107

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *Journal of Neuroscience*, *17*(11), 4302–4311. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997

Keating, C. T., Hickman, L., Leung, J., Monk, R., Montgomery, A., Heath, H., & Sowden, S. (2023). Autism-related language preferences of English-speaking individuals across the globe: A mixed methods investigation. *Autism Research*, *16*(2), 406–428. https://doi.org/10.1002/aur.2864

Klin, A., Danovitch, J. H., Merz, A. B., & Volkmar, F. R. (2007). Circumscribed Interests in Higher Functioning Individuals With Autism Spectrum Disorders: An Exploratory Study. *Research and Practice for Persons with Severe Disabilities*, *32*(2), 89–100. https://doi.org/10.2511/rpsd.32.2.89

Kohls, G., Antezana, L., Mosner, M. G., Schultz, R. T., & Yerys, B. E. (2018). Altered reward system reactivity for personalized circumscribed interests in autism. *Molecular Autism*, *9*(1). https://doi.org/10.1186/s13229-018-0195-7

Kosakowski, H. L., Cohen, M. A., Takahashi, A., Keil, B., Kanwisher, N., & Saxe, R. (2022). Selective responses to faces, scenes, and bodies in the ventral visual pathway of infants. *Current Biology*, *32*(2), 265-274.e5. https://doi.org/10.1016/j.cub.2021.10.064

Kosakowski, H. L., Norman-Haignere, S., Mynick, A., Takahashi, A., Saxe, R., & Kanwisher, N. (2023). Preliminary evidence for selective cortical responses to music in one-month-old infants. *Developmental Science*, *n/a*(n/a), e13387. https://doi.org/10.1111/desc.13387

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, *5*(11), Article 11. https://doi.org/10.1038/nrn1533

Kuhl, P. K. (2007). Is speech learning 'gated' by the social brain? *Developmental Science*, *10*(1), 110–120. https://doi.org/10.1111/j.1467-7687.2007.00572.x

Kuhl, P. K., Tsao, F.-M., & Liu, H.-M. (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences*, *100*(15), 9096–9101. https://doi.org/10.1073/pnas.1532872100

Law, J., & Roy, P. (2008). Parental Report of Infant Language Skills: A Review of the Development and Application of the Communicative Development Inventories. *Child and Adolescent Mental Health*, *13*(4), 198–206. https://doi.org/10.1111/j.1475-3588.2008.00503.x

Lidzba, K., Schwilling, E., Grodd, W., Krägeloh-Mann, I., & Wilke, M. (2011). Language comprehension vs. language production: Age effects on fMRI activation. *Brain and Language*, *119*(1), 6–15. https://doi.org/10.1016/j.bandl.2011.02.003

Liu, Y.-F., Kim, J., Wilson, C., & Bedny, M. (2020). Computer code comprehension shares neural resources with formal logical inference in the fronto-parietal network. *ELife*, *9*, e59340. https://doi.org/10.7554/eLife.59340

MacSweeney, M., Capek, C. M., Campbell, R., & Woll, B. (2008). The signing brain: The neurobiology of sign language. *Trends in Cognitive Sciences*, *12*(11), 432–440. https://doi.org/10.1016/j.tics.2008.07.010

Malik-Moraleda, S., Ayyash, D., Gallée, J., Affourtit, J., Hoffmann, M., Mineroff, Z., Jouravlev, O., & Fedorenko, E. (2022). An investigation across 45 languages and 12 language families reveals a universal language network. *Nature Neuroscience*, *25*(8), Article 8. https://doi.org/10.1038/s41593-022-01114-5

Menenti, L., Gierhan, S. M. E., Segaert, K., & Hagoort, P. (2011). Shared Language: Overlap and Segregation of the Neuronal Infrastructure for Speaking and Listening Revealed by Functional MRI. *Psychological Science*, *22*(9), 1173–1182. https://doi.org/10.1177/0956797611418347

Mitra, A., Snyder, A. Z., Tagliazucchi, E., Laufs, H., Elison, J., Emerson, R. W., Shen, M. D., Wolff, J. J., Botteron, K. N., Dager, S., Estes, A. M., Evans, A., Gerig, G., Hazlett, H. C., Paterson, S. J., Schultz, R. T., Styner, M. A., Zwaigenbaum, L., Network, T. I., … Raichle, M. (2017). Resting-state fMRI in sleeping infants more closely resembles adult sleep than adult wakefulness. *PLOS ONE*, *12*(11), e0188122. https://doi.org/10.1371/journal.pone.0188122

Mody, M., & Belliveau, J. W. (2013). Speech and Language Impairments in Autism: Insights from Behavior and Neuroimaging. *North American Journal of Medicine & Science*, *5*(3), 157–161.

Monti, M. M., Parsons, L. M., & Osherson, D. N. (2012). Thought Beyond Language: Neural Dissociation of Algebra and Natural Language. *Psychological Science*, *23*(8), 914–922. https://doi.org/10.1177/0956797612437427

Moore-Parks, E. N., Burns, E. L., Bazzill, R., Levy, S., Posada, V., & Müller, R.-A. (2010). An fMRI study of sentence-embedded lexical-semantic decision in children and adults. *Brain and Language*, *114*(2), 90–100. https://doi.org/10.1016/j.bandl.2010.03.009

Neville, H. J., Bavelier, D., Corina, D., Rauschecker, J., Karni, A., Lalwani, A., Braun, A., Clark, V., Jezzard, P., & Turner, R. (1998). Cerebral organization for language in deaf and hearing subjects: Biological constraints and effects of experience. *Proceedings of the National Academy of Sciences*, *95*(3), 922–929. https://doi.org/10.1073/pnas.95.3.922

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*(1), 11–28. https://doi.org/10.1016/0364-0213(90)90024-Q

Olulade, O. A., Seydell-Greenwald, A., Chambers, C. E., Turkeltaub, P. E., Dromerick, A. W., Berl, M. M., Gaillard, W. D., & Newport, E. L. (2020). The neural basis of language development: Changes in lateralization over age. *Proceedings of the National Academy of Sciences*, *117*(38), 23477–23483. https://doi.org/10.1073/pnas.1905590117

Onishi, K. H., & Baillargeon, R. (2005). Do 15-Month-Old Infants Understand False Beliefs? *Science*, *308*(5719), 255–258. https://doi.org/10.1126/science.1107621

Overath, T., McDermott, J. H., Zarate, J. M., & Poeppel, D. (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nature Neuroscience*, *18*(6), Article 6. https://doi.org/10.1038/nn.4021

Pan, B. A., Rowe, M. L., Spier, E., & Tamis-LeMonda, C. (2004). Measuring productive vocabulary of toddlers in low-income families: Concurrent and predictive validity of three sources of data. *Journal of Child Language*, *31*(3), 587–608. https://doi.org/10.1017/s0305000904006270

Paunov, A. M., Blank, I. A., & Fedorenko, E. (2019). Functionally distinct language and Theory of Mind networks are synchronized at rest and during language comprehension. *Journal of Neurophysiology*, *121*(4), 1244–1265. https://doi.org/10.1152/jn.00619.2018

Paunov, A. M., Blank, I. A., Jouravlev, O., Mineroff, Z., Gallée, J., & Fedorenko, E. (2022). Differential Tracking of Linguistic vs. Mental State Content in Naturalistic Stimuli by Language and Theory of Mind (ToM) Brain Networks. *Neurobiology of Language*, 1–29. https://doi.org/10.1162/nol_a_00071

Perani, D., Saccuman, M. C., Scifo, P., Anwander, A., Spada, D., Baldoli, C., Poloniato, A., Lohmann, G., & Friederici, A. D. (2011). Neural language networks at birth. *Proceedings of the National Academy of Sciences*, *108*(38), 16056–16061. https://doi.org/10.1073/pnas.1102991108

Perani, D., Saccuman, M. C., Scifo, P., Spada, D., Andreolli, G., Rovelli, R., Baldoli, C., & Koelsch, S. (2010). Functional specializations for music processing in the human newborn brain. *Proceedings of the National Academy of Sciences*, *107*(10), 4758–4763. https://doi.org/10.1073/pnas.0909074107

Pickles, A., Anderson, D. K., & Lord, C. (2014). Heterogeneity and plasticity in the development of language: A 17-year follow-up of children referred early for possible autism. *Journal of Child Psychology and Psychiatry*, *55*(12), 1354–1362. https://doi.org/10.1111/jcpp.12269

Pierce, K., & Redcay, E. (2008). Fusiform Function in Children with an Autism Spectrum Disorder Is a Matter of "Who." *Biological Psychiatry*, *64*(7), 552–560. https://doi.org/10.1016/j.biopsych.2008.05.013

Poldrack, R. A., Paré-Blagoev, E. J., & Grant, P. E. (2002). Pediatric Functional Magnetic Resonance Imaging: Progress and Challenges. *Topics in Magnetic Resonance Imaging*, *13*(1), 61.

Price, C. J. (2010). The anatomy of language: A review of 100 fMRI studies published in 2009. *Annals of the New York Academy of Sciences*, *1191*(1), 62–88. https://doi.org/10.1111/j.1749-6632.2010.05444.x

Raschle, N., Zuk, J., Ortiz-Mantilla, S., Sliva, D. D., Franceschi, A., Grant, P. E., Benasich, A. A., & Gaab, N. (2012). Pediatric neuroimaging in early childhood and infancy: Challenges and practical guidelines. *Annals of the New York Academy of Sciences*, *1252*(1), 43–50. https://doi.org/10.1111/j.1749-6632.2012.06457.x

Redcay, E., & Courchesne, E. (2008). Deviant Functional Magnetic Resonance Imaging Patterns of Brain Activity to Speech in 2–3-Year-Old Children with Autism Spectrum Disorder. *Biological Psychiatry*, *64*(7), 589–598. https://doi.org/10.1016/j.biopsych.2008.05.020

Redcay, E., Haist, F., & Courchesne, E. (2008). Functional neuroimaging of speech perception during a pivotal period in language acquisition. *Developmental Science*, *11*(2), 237–252. https://doi.org/10.1111/j.1467-7687.2008.00674.x

Redcay, E., & Moraczewski, D. (2020). Social cognition in context: A naturalistic imaging approach. *NeuroImage*, *216*, 116392. https://doi.org/10.1016/j.neuroimage.2019.116392

Restrepo, M. A., Schwanenflugel, P. J., Blake, J., Neuharth-Pritchett, S., Cramer, S. E., & Ruston, H. P. (2006). Performance on the PPVT–III and the EVT: Applicability of the Measures With African American and European American Preschool Children. *Language, Speech, and Hearing Services in Schools*, *37*(1), 17–27. https://doi.org/10.1044/0161-1461(2006/003)

Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., & Saxe, R. (2018). Development of the social brain from age three to twelve years. *Nature Communications*, *9*(1), Article 1. https://doi.org/10.1038/s41467-018-03399-2

Robison, J. E. (2019). Talking about autism—Thoughts for researchers. *Autism Research*, *12*(7), 1004–1006. https://doi.org/10.1002/aur.2119

Romeo, R. R., Leonard, J. A., Robinson, S. T., West, M. R., Mackey, A. P., Rowe, M. L., & Gabrieli, J. D. E. (2018). Beyond the 30-Million-Word Gap: Children's Conversational Exposure Is Associated With Language-Related Brain Function. *Psychological Science*, *29*(5), 700–710. https://doi.org/10.1177/0956797617742725

Rowe, M. L. (2008). Child-directed speech: Relation to socioeconomic status, knowledge of child development and child vocabulary skill*. *Journal of Child Language*, *35*(1), 185–205. https://doi.org/10.1017/S0305000907008343

Satterthwaite, T. D., Wolf, D. H., Loughead, J., Ruparel, K., Elliott, M. A., Hakonarson, H., Gur, R. C., & Gur, R. E. (2012). Impact of in-scanner head motion on multiple measures of functional connectivity: Relevance for studies of neurodevelopment in youth. *NeuroImage*, *60*(1), 623–632. https://doi.org/10.1016/j.neuroimage.2011.12.063

Saxe, R., Brett, M., & Kanwisher, N. (2006). Divide and conquer: A defense of functional localizers. *NeuroImage*, *30*(4), 1088–1096. https://doi.org/10.1016/j.neuroimage.2005.12.062

Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding Other Minds: Linking Developmental Psychology and Functional Neuroimaging. *Annual Review of Psychology*, *55*, 87–124. https://doi.org/10.1146/annurev.psych.55.090902.142044

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking peopleThe role of the temporo-parietal junction in "theory of mind." *NeuroImage*, *19*(4), 1835–1842. https://doi.org/10.1016/S1053-8119(03)00230-1

Schlosser, M. J., Aoyagi, N., Fulbright, R. K., Gore, J. C., & McCarthy, G. (1998). Functional MRI studies of auditory comprehension. *Human Brain Mapping*, *6*(1), 1–13. https://doi.org/10.1002/(SICI)1097-0193(1998)6:1<1::AID-HBM1>3.0.CO;2-7

Scott, T. L., Gallée, J., & Fedorenko, E. (2017). A new fun and robust version of an fMRI localizer for the frontotemporal language system. *Cognitive Neuroscience*, *8*(3), 167–176. https://doi.org/10.1080/17588928.2016.1201466

Shain, C., Paunov, A., Chen, X., Lipkin, B., & Fedorenko, E. (2022). *No evidence of theory of mind reasoning in the human language network* (p. 2022.07.18.500516). bioRxiv. https://doi.org/10.1101/2022.07.18.500516

Shaywitz, S. E., Shaywitz, B. A., Pugh, K. R., Fulbright, R. K., Constable, R. T., Mencl, W. E., Shankweiler, D. P., Liberman, A. M., Skudlarski, P., Fletcher, J. M., Katz, L., Marchione, K. E., Lacadie, C., Gatenby, C., & Gore, J. C. (1998). Functional disruption in the organization of the brain for reading in dyslexia. *Proceedings of*

the National Academy of Sciences, *95*(5), 2636–2641. https://doi.org/10.1073/pnas.95.5.2636

Shnayer, S. W. (1968). *Some Relationships between Reading Interest and Reading Comprehension*. https://eric.ed.gov/?id=ED022633

Snowling, M. J. (2001). From language to reading and dyslexia1. *Dyslexia*, *7*(1), 37–46. https://doi.org/10.1002/dys.185

Sroka, M. C., Vannest, J., Maloney, T. C., Horowitz-Kraus, T., Byars, A. W., Holland, S. K., & CMIND Authorship Consortium. (2015). Relationship between receptive vocabulary and the neural substrates for story processing in preschoolers. *Brain Imaging and Behavior*, *9*(1), 43–55. https://doi.org/10.1007/s11682-014-9342-8

Stoppelman, N., Harpaz, T., & Ben-Shachar, M. (2013). Do not throw out the baby with the bath water: Choosing an effective baseline for a functional localizer of speech processing. *Brain and Behavior*, *3*(3), 211–222. https://doi.org/10.1002/brb3.129

Szaflarski, J. P., Altaye, M., Rajagopal, A., Eaton, K., Meng, X., Plante, E., & Holland, S. K. (2012). A 10-year longitudinal fMRI study of narrative comprehension in children and adolescents. *NeuroImage*, *63*(3), 1188–1195. https://doi.org/10.1016/j.neuroimage.2012.08.049

Szaflarski, J. P., Holland, S. K., Schmithorst, V. J., & Byars, A. W. (2006). FMRI study of language lateralization in children and adults. *Human Brain Mapping*, *27*(3), 202–212. https://doi.org/10.1002/hbm.20177

ten Cate, C., & Okanoya, K. (2012). Revisiting the syntactic abilities of non-human animals: Natural vocalizations and artificial grammar learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1598), 1984–1994. https://doi.org/10.1098/rstb.2012.0055

Tomova, L., Wang, K. L., Thompson, T., Matthews, G. A., Takahashi, A., Tye, K. M., & Saxe, R. (2020). Acute social isolation evokes midbrain craving responses similar to hunger. *Nature Neuroscience*, *23*(12), Article 12. https://doi.org/10.1038/s41593-020-00742-z

Tuckute, G., Paunov, A., Kean, H., Small, H., Mineroff, Z., Blank, I., & Fedorenko, E. (2022). Frontal language areas do not emerge in the absence of temporal language areas: A case study of an individual born without a left temporal lobe. *Neuropsychologia*, *169*, 108184. https://doi.org/10.1016/j.neuropsychologia.2022.108184

Vanderwal, T., Eilbott, J., & Castellanos, F. X. (2019). Movies in the magnet: Naturalistic paradigms in developmental functional neuroimaging. *Developmental Cognitive Neuroscience*, *36*, 100600. https://doi.org/10.1016/j.dcn.2018.10.004

Vanderwal, T., Kelly, C., Eilbott, J., Mayes, L. C., & Castellanos, F. X. (2015). Inscapes: A movie paradigm to improve compliance in functional magnetic resonance imaging. *NeuroImage*, *122*, 222–232. https://doi.org/10.1016/j.neuroimage.2015.07.069

Walker, D., Greenwood, C., Hart, B., & Carta, J. (1994). Prediction of School Outcomes Based on Early Language Production and Socioeconomic Factors. *Child Development*, *65*(2), 606–621. https://doi.org/10.1111/j.1467-8624.1994.tb00771.x

Wehbe, L., Blank, I. A., Shain, C., Futrell, R., Levy, R., von der Malsburg, T., Smith, N., Gibson, E., & Fedorenko, E. (2021). Incremental Language Comprehension Difficulty Predicts Activity in the Language Network but Not the Multiple Demand Network. *Cerebral Cortex*, *31*(9), 4006–4023. https://doi.org/10.1093/cercor/bhab065

Wernicke, C. (1874). *Der aphasische Symptomencomplex: Eine psychologische Studie auf anatomischer Basis*. Cohn & Weigert.

Wild, C. J., Linke, A. C., Zubiaurre-Elorza, L., Herzmann, C., Duffy, H., Han, V. K., Lee, D. S. C., & Cusack, R. (2017). Adult-like processing of naturalistic sounds in auditory cortex by 3- and 9-month old infants. *NeuroImage*, *157*, 623–634. https://doi.org/10.1016/j.neuroimage.2017.06.038

Wood, A. G., Harvey, A. S., Wellard, R. M., Abbott, D. F., Anderson, V., Kean, M., Saling, M. M., & Jackson, G. D. (2004). Language cortex activation in normal children. *Neurology*, *63*(6), 1035–1044. https://doi.org/10.1212/01.WNL.0000140707.61952.CA

Yates, T. S., Ellis, C. T., & Turk-Browne, N. B. (2021a). Emergence and organization of adult brain function throughout child development. *NeuroImage*, *226*, 117606. https://doi.org/10.1016/j.neuroimage.2020.117606

Yates, T. S., Ellis, C. T., & Turk-Browne, N. B. (2021b). The promise of awake behaving infant fMRI as a deep measure of cognition. *Current Opinion in Behavioral Sciences*, *40*, 5–11. https://doi.org/10.1016/j.cobeha.2020.11.007

"Using words to talk of words is like using a pencil to draw a picture of itself, on itself. Impossible. Confusing. Frustrating ... but there are other ways to understanding."

— Patrick Rothfuss, ***The Name of the Wind***

Chapter 2 : Left-hemisphere cortical language regions respond equally to dialogue and monologue

**A version of this chapter has been submitted for publication as:*

Olson, H. A., Chen, E. M., Lydic, K. O., & Saxe, R. R. Left-hemisphere cortical language regions respond equally to dialogue and monologue.
Preprint: *https://www.biorxiv.org/content/10.1101/2023.01.30.526344v1*

Abstract

Much of the language we encounter in our everyday lives comes in the form of conversation, yet the majority of research on the neural basis of language comprehension has used language input from a single source. To determine whether canonical left-hemisphere language regions are sensitive to features of dialogue beyond the comprehensibility of the speech stream, we scanned 20 adults on two novel tasks using functional magnetic resonance imaging. In the first, participants watched videos of puppets speaking either to the viewer (monologue) or to a partner (dialogue), while the audio was either comprehensible (forward) or reversed (backward). Canonical left-hemisphere language regions responded more to forward than backward speech, as expected, but did not respond more to dialogue than monologue. In a second task, two puppets conversed with each other, but only one was comprehensible while the other's speech stream was reversed. Left-hemisphere cortical language regions again responded more to forward than backwards speech, and activity in these regions was only correlated among participants who heard the same characters speaking forward and backward. In contrast, some theory of mind regions and right hemisphere homologues of language regions responded more to dialogue than monologue, and activity in some of these regions was correlated among

participants even when opposite characters were speaking forward and backward (in both cases, the visual video clips were held constant). Together, these experiments suggest that canonical left-hemisphere cortical language regions are only sensitive to the language input in dialogue.

Introduction

Language is first heard, learned and used in informal conversation. By contrast, most research on the neural basis of language comprehension has relied on language from a single source. Compared to monologues or single-source texts, language in turn-taking dialogue exhibits distinctive features that function to coordinate and monitor the creation of common ground (Clark, 1996; Clark & Schaefer, 1989; Fox Tree, 1999; Fusaroli & Tylén, 2016). Successive utterances not only convey new meaning, but often show how a prior utterance was understood, facilitating rapid correction (Schegloff et al., 1977). In conversation, speakers quickly volley back and forth, establishing referents across speaker boundaries and often finishing each other's sentences (Clark, 1996; Clark & Schaefer, 1989; Clark & Wilkes-Gibbs, 1986). Speakers alternate about every 2 seconds, with only a 200 ms delay between their utterances on average (Levinson, 2016; Stivers et al., 2009). When observing conversation, adults and even young children can accurately predict turn taking (Casillas & Frank, 2017). Consequently, although utterances in dialogue are typically not well-formed grammatical sentences, dialogue is easier to comprehend than monologue from a single speaker (Fox Tree, 1999).

Here, for example, is a short transcribed excerpt from a two-speaker dialogue:

> *Well, you see, I've never met him, and so if he comes to the door, how will I know that it's him? Ah. Oh well, it's easy. For one thing, we're exactly alike. You are? Yeah! We're twins!*[8]

As a single linguistic stream, this excerpt is hard to understand, including sentence fragments and seeming disfluencies. Yet when the utterances are assigned to different speakers, the dialogue becomes easily comprehensible:

> ***Ernie**: Well, you see, I've never met him, and so if he comes to the door, how will I know that it's him?*
> ***Bert**: Ah. Oh well, it's easy. For one thing, we're exactly alike.*
> ***Ernie**: You are?*
> ***Bert**: Yeah! We're twins!*

In this example, the backchannel utterance "*Ah*" conveys understanding, and the clarifying question "*You are?*" marks distinct perspectives. Rather than seeming disfluent, these utterances aid in language comprehension (Clark, 1996; Clark & Schaefer, 1989; Tolins & Fox Tree, 2016). Bert knows something about the absent referent that Ernie does not (i.e., that this person is Bert's twin). Representing these different perspectives is integral to understanding the dialogue and predicting what might come next. Comprehending language in dialogue thus requires additional social and linguistic processing, compared to comprehending language from a single source.

[8] Source: https://youtu.be/sS7_-h882Ls

Because of this feature, observed dialogue provides an interesting test case for probing the functions of cortical regions involved in language processing. A consistent set of left hemisphere frontal and temporal regions are involved in processing language (Bates et al., 2001; Binder et al., 1997; Dronkers et al., 2004; Fedorenko et al., 2010, 2011; Friederici, 2011; Friederici & Gierhan, 2013; Price, 2010, 2012). These regions, which we will refer to as the 'canonical language network', robustly respond to language whether it is spoken (Scott et al., 2017), written (Fedorenko et al., 2010), or signed (MacSweeney et al., 2008; Neville et al., 1998). They are active during both production and comprehension (Hagoort, 2014; Hu et al., 2022; Menenti et al., 2011; Price, 2010), in adults and children (Enge et al., 2020), across a wide range of languages (Malik-Moraleda et al., 2022). They are also sensitive to features of language like comprehension difficulty (Wehbe et al., 2021) and syntactic complexity (Blank et al., 2016), responding more to higher syntactic and semantic processing demands (Hagoort & Indefrey, 2014).

Since early lesions studies, it has generally been agreed that canonical language network regions are *necessary* for language (Broca, 1865; Wernicke, 1874). However, there have been long standing debates about the *specificity* of these regions for language processing, and what their limits and scope may be (Fedorenko & Thompson-Schill, 2014; Monti et al., 2012). Initially, whole brain activation mapping suggested that language activates regions that overlapped with a range of other cognitive tasks (Blumstein & Amso, 2013; Gold & Buckner, 2002; Thompson-Schill et al., 1997). When language regions are functionally localized within individuals (Braga et al., 2020; Fedorenko et al., 2010), these regions are not engaged by other compositional or cognitively difficult tasks, like working memory, math, music, cognitive control, action observation, or imitation (Fedorenko et al., 2011; Pritchett et

al., 2018). But what about a task that is more cognitively similar? Recent studies looked at a particularly interesting boundary case: reading computer code. Coding shares a number of features with language processing: both, for instance, involve recursively combining components in constrained ways to form a more complex meaning (Fedorenko et al., 2019). Despite the underlying similarities, language regions are not recruited when people read and evaluate the meaning of computer code (Ivanova et al., 2020; Liu et al., 2020), providing further evidence that language regions are highly specific to language processing.

Observed dialogue is another interesting boundary case for probing the scope of language regions because listening to dialogue requires the same linguistic processes as listening to monologue, plus additional processes involved in tracking the alternations between speakers. This additional processing may or may not rely on canonical language regions. For instance, compared to processing linguistic input from a single speaker, understanding overheard dialogue requires tracking the difference between at least two speakers' perspectives; thus, understanding dialogue may rely more on Theory of Mind (ToM) - our ability to reason about others' minds - than understanding monologue. ToM selectivity engages a network of regions in right and left temporoparietal junction (RTPJ, LTPJ), middle, ventral, and dorsal parts of medial prefrontal cortex (MMPFC, VMPFC, DMPFC), and precuneus (PC) (Dufour et al., 2013; Saxe & Kanwisher, 2003; Saxe & Powell, 2006). Activity in networks for language and ToM is correlated when processing socially relevant language (Paunov et al., 2019), but these networks still maintain their functional specificity and distribution of roles (Paunov et al., 2022). When comprehending dialogue, tracking shifts in speaker perspective may also specifically be relevant for language processing, as the speaker's perspective

can impact the interpretation of the utterance and the predictability of the subsequent response.

Prior research has looked at the neural correlates of comprehension in dialogue when the meaning of an utterance depends on the preceding utterance and contextual information. For example, the utterance "it's hard to give a good presentation" could be a direct response to the question "how difficult is it to prepare a presentation?" (difficult), or an indirect response to the question "what did you think of my presentation?" (not so great; examples adapted from Bašnáková et al., 2014). In the brain, ToM network regions including DMPFC and RTPJ, as well as bilateral IFG, and right MTG, responded more to the same utterance when it was an indirect response than when it was a direct responses (Bašnáková et al., 2014; Feng et al., 2017). Another study found that left temporal and frontal regions responded more to indirect than direct replies in question-response pairs (Jang et al., 2013), but the sentences were not controlled for linguistic features between conditions, unlike Bašnáková et al and Feng et al. Individuals with high communicative skills also showed more activation than individuals with low communicative skills for indirect>direct responses in dialogue in regions outside either language or ToM network (Bendtz et al., 2022). These results suggest that the processing of implied meaning in indirect responses mostly occurs outside of the core language network. However, this conclusion remains uncertain. Activation near IFG might imply modulation of the core language network, because part of left IFG is in the canonical language network. This pattern could also reflect activation of nearby 'multiple demand' regions that respond to task difficulty (Blank et al., 2014; Fedorenko et al., 2012; Fedorenko & Blank, 2020), as the indirect replies elicited slower reaction times than the direct replies (Feng et al., 2017) and these studies did not use subject-specific functional regions of interest (ss-fROIs). As

experimental stimuli, auditory question-response pairs are well controlled, but afford limited opportunity to recognize, mark and then resolve differences of perspectives between speakers.

Thus, it remains an open question whether the processes that enable an observer to track the alternating perspectives between interlocutors, which are integral to dialogue comprehension, lie within the scope of canonical language regions. To probe the sensitivity of canonical language regions to dialogue, we first directly compared activity in individuals' canonical language regions during naturalistic, audiovisual excerpts of monologue and dialogue. We chose videos in order to maximize meaningful tracking of speakers as individuals with distinct perspectives; consequently, we also needed to create a control condition that accounted for the visual differences between dialogue and monologue. To do this, we used a block-design task ("Sesame Street-Blocked Language" or SS-BlockedLang) with conditions controlling for both comprehensible speech and conversational interaction: videos of two characters (from *Sesame Street*) engaging in either a dialogue or monologues, with the audio for each utterance played normally or temporally reversed (Experiment 1). We expected language regions to respond more to comprehensible than incomprehensible speech, but critically, we tested whether language regions responded differently to dialogue versus monologue. A difference in activity could go in either direction: because dialogue requires tracking multiple speakers' speech streams, activity could be higher in dialogue than monologue; however, behaviorally, dialogue is easier to understand than monologue (Fox Tree, 1999; Garrod & Pickering, 2004), and thus may evoke less activity in language regions. To identify language regions in each individual, we used an independent functional localizer task (Scott et al., 2017).

While Experiment 1 was designed to detect differences in experimenter-manipulated aspects of the language clips (dialogue versus monologue, comprehensible versus incomprehensible), one concern with only using an experimentally-controlled block design is that it might be insensitive to unspecified sources of continuous variation in the neural response (Hamilton & Huth, 2020; Hasson et al., 2004; Hasson & Honey, 2012; Nastase et al., 2020). Complementary efforts within the field of neuroscience have used naturalistic stimuli that are more ecologically valid, such as movies or narratives (Grall & Finn, 2022; Hamilton & Huth, 2020; Nastase et al., 2021; Sonkusare et al., 2019). Therefore, we also designed a complementary second experiment ("Sesame Street-Interleaved Dialogue" or SS-IntDialog) involving longer (1-3 minute) continuous clips of dialogue between two characters (Experiment 2). Rather than manipulating forwards vs backwards speech per video clip, as in Experiment 1, in these longer clips, we reversed one of the two character's utterances, such that one character spoke forwards and the other spoke backwards. The visual input was the same for all participants, but the auditory input was not: which character spoke in forward versus backwards speech, in each video, was flipped for half of the participants. First, we tested whether language regions still responded robustly to comprehensible speech, even though the surrounding language (e.g., the speech stream of the other interlocutor) was rendered incomprehensible, meaning that the comprehensible utterances were highly variable and sometimes quite short. Next, we used inter-subject correlation (ISC) analysis to extract stimulus-driven variation in the naturalistic conversation clips, specifically to determine whether language regions only tracked the language-driven variation, or whether they also tracked other aspects of the dialogue that were conveyed independent of which character was comprehensible in the clips (e.g., that two characters alternated speaking, that two characters were present, what objects were in the scene, what the scene was about, etc.). This allowed us to compare

the timecourses of response in participants who heard the same version of the stimuli (e.g., the same characters forward vs. backward) to participants who heard the flipped version of the stimuli (e.g., opposite characters forward vs. backward). If language regions are engaged only in the processing of comprehensible language, and not are not involved in processing any other information that is conveyed in the dialogue clips, then we would expect no correlation in responses across participants who heard flipped versions of the same video.

Regardless of whether canonical language regions differentiate between dialogue and monologue, other regions may show a sensitivity to dialogue. We tested two specific possibilities using individually-defined ss-fROIs: ToM regions and right hemisphere homologues of language regions. Given that speaker alternations in dialogue require integrating across distinct perspectives, we hypothesized that ToM regions might respond differently to dialogue than monologue. Right hemisphere damage can make it more difficult for individuals to make inferences from discourse (Beeman, 1993), and prior work has demonstrated the right hemisphere's preferential involvement in social and contextual aspects of language processing (Friederici, 2011; Frühholz et al., 2012; Ross & Monnot, 2008; Seydell-Greenwald et al., 2020); thus, it was also possible that right hemisphere homologues of language regions might be sensitive to features of dialogue other than just comprehensibility of language. A third possibility is that other regions may be specifically involved in processing comprehensible dialogue, such as regions involved in processing social interactions (Isik et al., 2017); thus, we also performed a whole-brain analysis.

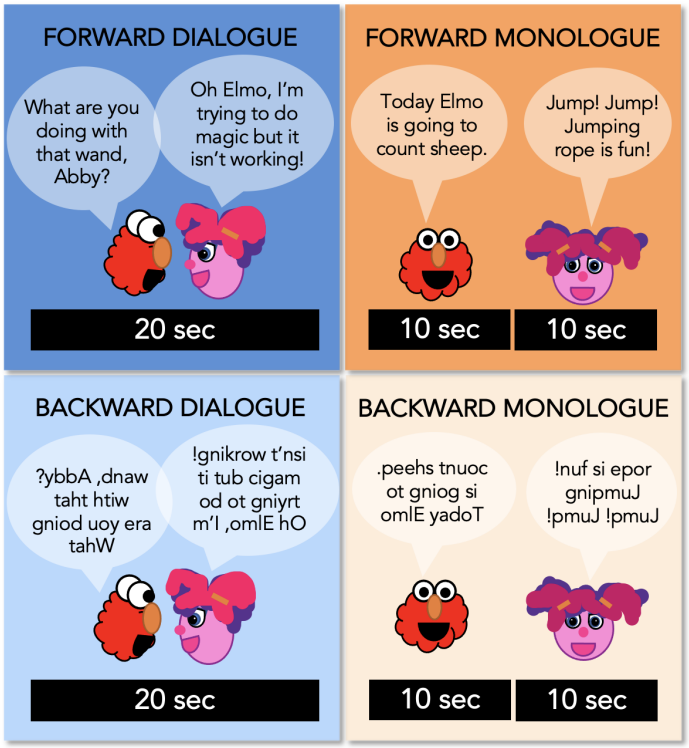
Experiment 1: SS-BlockedLang

In Experiment 1, we ran our novel 2x2 block localizer (SS-BlockedLang) along with standard pre-existing localizers for language (Scott et al., 2017) and ToM (Dodell-Feder et al., 2011) to measure responses to dialogue and monologue speech in subject-specific functional regions of interest for language and ToM.

Methods

Stimuli Design: Our goal was to create a set of stimuli that allowed us to manipulate both comprehensibility and dialogue vs. monologue in a 2x2 block task design (**Figure 2.1**). To do this, we created a set of 20-second edited audiovisual clips of *Sesame Street* during which either two puppets speak to each other (Dialogue), or a single puppet addresses the viewer (Monologue), while the auditory speech stream is played either normally (Forward) or reversed so as to be incomprehensible (Backward). Dialogue blocks consisted of two characters, both present in the same scene, speaking back-and-forth for a total of 20 seconds, and Monologue blocks consisted of two sequential 10-second clips of a character, present alone. In the Backward conditions, the audio was reversed within each character rather than across the entire clip, ensuring a continuity of voice-character alignment. For instance, in a Backward Dialogue block with Elmo and Abby, Elmo's voice was reversed and played when Elmo was talking, and Abby's voice was reversed and played while Abby was talking.

Figure 2.1: SS-BlockedLang Task Design.



Participants watched 20-second clips of Dialogue (blue) and Monologue (orange) of Sesame Street, in which the audio is played either Forward or Backward.

We chose to use audiovisual stimuli in order to increase participant engagement with the stimuli, facilitate language comprehension, and emphasize the context of the dialogue by showing multiple characters interacting on the screen. However, using audiovisual stimuli rather than audio-only stimuli introduced a challenge: how do we avoid creating distracting cross-modal mismatches while only varying the auditory, and not the visual, input across conditions? Even infants and young children are sensitive to the congruence between a speaker's mouth movements and the sounds they produce in speech (Gogate & Bahrick, 1998; Lewkowicz & Flom, 2014). To balance these desiderata, we decided to use puppets with rigid mouths (rather than human actors) so that the congruence between mouth movements and audio would be more similar

between the forward and backward speech conditions. Thus, we could counterbalance whether the audio was played forward or backward across participants while showing them the exact same visual clips.

A notable feature of our task is that it uses commercially produced video clips that were not designed for research purposes. Because we intended to eventually use these same stimuli with very young children, video clips were selected from episodes of *Sesame Street* to appeal to a wide age range. The linguistic content is embedded within colorful, dynamic videos with different characters, different voices, and different settings. To retain the natural feel of the clips, the audio was reversed within each utterance of a particular character and carefully overlaid such that the reversed audio still reasonably matched up with the puppets' mouth movements, and each character's "voice" was still unique when the audio was reversed. To create the stimuli, we adhered to the following guidelines: (1) we only selected clips that had an overall neutral or positive valence, (2) we only included clips of puppets, rather than clips with humans and puppets, (3) we excluded clips in which the reversed speech did not align well with mouth movements, and (4) we left non-linguistic sounds in the clips, aiming to retain the integrity of the content. Transcripts and stimuli features can be found on OSF[9].

Because we selected commercially available clips, we did not control for linguistic properties of the stimuli. Monologue and dialogue blocks did not differ in the number of mental state words per block or the total number of words per block. However,

[9] https://osf.io/whsb7/

monologue blocks had significantly longer mean length of utterance, and a lower Flesh-Kincaid reading ease score (see **Chapter 3** for details).

Preregistration: Methods and hypotheses were preregistered on OSF: validation as language localizer[10] and analyses of conversation processing[11]. We did not preregister any hypotheses about right hemisphere homologues of language regions. There were a few deviations from the initial preregistrations for methods: (1) We used a different version of fMRIprep than specified in the preregistration. (2) We decided not to exclude runs of the ToM localizer task based on performance on the true/false questions. (3) We did not preregister testing for effects at the network level, but decided to include these tests along with effects at the ROI level. (4) Not all of the analyses specified in the preregistrations are included in this paper, including: analyses involving laterality index, analyses varying the significance threshold (we used our primary preregistered threshold of $p<.001$), analyses directly comparing the effect of matched vs. mismatched audio in SS-IntDialog in language vs. ToM regions, whole-brain analyses of forward monologue-specific effects, and analyses comparing variance explained in language regions' response using the experimenter-derived regressor of forward/backward speech vs. average timecourse of within-group subjects in the SS-IntDialog task.

Participants: We scanned 20 adults (age: mean(SD) = 23.85(3.70) years, range 18-30 years) who were fluent speakers of English, right-handed, and had no MRI contraindications. Recruitment was restricted to adults with access to the MIT campus

[10] https://osf.io/n4ur5/
[11] https://osf.io/kzdpc/

according to Covid-19 policies. The protocol was approved by the MIT Committee on the Use of Humans as Experimental Subjects. Informed consent was provided by all participants. Participants were compensated at a rate of $30/hour for scanning, which is standard for our lab and imaging center.

Experimental Protocol: Data were acquired from a 3-Tesla Siemens Magnetom Prisma scanner located at the Athinoula A. Martinos Imaging Center at MIT using a 32-channel head coil. The scanning session lasted approximately 90 minutes and included a T1 anatomical scan and 10 functional scans: 4 runs of SS-BlockedLang (Exp. 1), 2 runs of SS-IntDialog (Exp. 2), 2 runs of the auditory language localizer (Scott et al., 2017), and 2 runs of the theory of mind localizer (Dodell-Feder et al., 2011). T1-weighted structural images were acquired in 176 interleaved sagittal slices with 1.0mm isotropic voxels (MPRAGE; TA=5:53; TR=2530.0ms; FOV=256mm; GRAPPA parallel imaging, acceleration factor of 2). Functional data were acquired with a gradient-echo EPI sequence sensitive to Blood Oxygenation Level Dependent (BOLD) contrast in 3mm isotropic voxels in 46 interleaved near-axial slices covering the whole brain (EPI factor=70; TR=2s; TE=30.0ms; flip angle=90 degrees; FOV=210mm). 185 volumes were acquired per run for SS-BlockedLang (TA=6:18), 262 volumes were acquired per run for SS-IntDialog (TA=8:52), 179 volumes were acquired per run for the auditory language localizer (TA=6:06), and 136 volumes were acquired per run for the ToM localizer (TA=4:40). fMRI tasks were run from a MacBook Pro laptop and projected onto a 16”x12” screen. Participants viewed the stimuli through a mirror attached to the head coil. Isocenter to screen + mirror to eye was 42" + 6" for both eyes. SS-BlockedLang and SS-IntDialog tasks were run through PsychoPy3 software version 3.2.4. The auditory language localizer and ToM localizer tasks were run through MATLAB version R2019a and PsychToolbox version 3.0.17.

fMRI Tasks:

SS-BlockedLang Language Task

We used a 2x2 block task design with four conditions: Forward Dialogue, Forward Monologue, Backward Dialogue, and Backward Monologue (Figure 2.1). Participants were asked to watch the 20-second videos and press a button on an in-scanner button box when they saw a still image of Elmo appear on the screen after each 20-second block. Participants completed 4 runs, each 6 min 18 sec long. Each run contained unique clips, and participants never saw a Forward and Backward version of the same clip. Each run contained 3 sets of 4 blocks, one of each condition (total of 12 blocks), with 22-second rest blocks after each set of 4 blocks. Forward and Backward versions of each clip were counterbalanced between participants (randomly assigned Set A or Set B). Run order was randomized for each participant.

Auditory Language Localizer

We used a task previously validated for identifying high-level language processing regions (Scott et al., 2017). Participants listened to Intact and Degraded 18-second blocks of speech. The Intact condition consisted of audio clips of spoken English (e.g., clips from interviews in which one person is speaking), and the Degraded condition consisted of acoustically degraded versions of these clips. Participants viewed a black dot on a white background during the task while passively listening to the auditory stimuli. 14-second fixation blocks (no sound) were present after every 4 speech blocks, as well as at the beginning and end of each run (5 fixation blocks per run). Participants completed two runs, each approximately 6 min 6 sec long. Each run contained 16 blocks of speech (8 intact, 8 degraded).

Theory of Mind Localizer

We used a task previously validated for identifying regions that are involved in ToM and social cognition (Dodell-Feder et al., 2011). Participants read short stories in two conditions: False Beliefs and False Photos. Stories in the False Beliefs condition described scenarios in which a character holds a false belief. Stories in the False Photos condition described outdated photographs and maps. Each story was displayed in white text on a black screen for 10 seconds, followed by a 4-second true/false question based on the story (which participants responded to via the in-scanner button box), followed by 12 seconds of a blank screen (rest). Each run contained 10 blocks. Participants completed two runs, each approximately 4 min 40 sec long.

fMRI Preprocessing and Statistical Modeling: FMRI data were first preprocessed using fMRIPrep 1.2.6 (Esteban et al., 2019), which is based on Nipype 1.1.7 (Gorgolewski et al., 2011). See Supplementary Materials for full preprocessing pipeline details. We used a lab-specific script that uses Nipype to combine tools from several different software packages for first-level modeling. Each event regressor was defined as a boxcar convolved with a standard double-gamma HRF, and a high-pass filter (1/210 Hz) was applied to both the data and the model. Artifact detection was performed using Nipype's RapidART toolbox (an implementation of SPM's ART toolbox). Individual TRs were marked as outliers if (1) there was more than .4 units of frame displacement, or (2) the average signal intensity of that volume was more than 3 standard deviations away from the mean average signal intensity. We included one regressor per outlier volume. In addition, we included a summary movement regressor (framewise displacement) and 6 anatomical CompCor regressors to control for the average signal in white matter and CSF. We applied a 6mm smoothing kernel to preprocessed BOLD images. The first-level model was run using FSL's GLM in MNI space. Subject level modeling was

performed with in-lab scripts using Nipype. Specifically, FSL's fixed effects flow was used to combine runs at the level of individual participants. A subject level model was created for each set of usable runs per contrast for each task (up to 4 runs for SS-BlockedLang, and up to 2 runs for SS-IntDialog, LangLoc, and ToMLoc). Runs with more than 20% of timepoints marked as outliers were excluded from analysis (1 run of SS-IntDialog in 1 participant and 1 run of ToMLoc in another participant were excluded for motion). We also excluded 1 run of SS-BlockedLang and 1 run of SS-IntDialog from a participant who reported falling asleep. Output average magnitudes in each voxel in the second level model were then passed to the group level model. Group modeling used in-lab scripts that implemented FSL's RANDOMISE to perform a nonparametric one-sample t-test of the con values against 0 (5000 permutations, MNI space, threshold alpha = .05), accounting for familywise error.

Group Whole Brain Analysis: For group whole brain results, we used a threshold of $p<.001$, corrected via threshold free cluster enhancement (TFCE corrected). For language comprehension, we used the [Forward Dialogue + Forward Monologue] > [Backward Dialogue + Backward Monologue] contrast. To determine whether any other regions in the brain are particularly responsive to comprehensible dialogue, we performed whole-brain analyses using the [Forward Dialogue > Forward Monologue] > [Backward Dialogue > Backward Monologue] contrast. In exploratory analyses, we used an uncorrected threshold of $p<.001$ (two-tailed, 19 degrees of freedom).

Individual Region of Interest Analysis for Language and Theory of Mind: We defined subject-specific functional regions of interest (ss-fROIs) for language as the top 100 voxels activated in an individual, within each of six predefined language search spaces, for the Intact>Degraded contrast using the auditory language localizer task (Fedorenko

et al., 2010). The six language search spaces in the left hemisphere included: Left IFGorb, Left IFG, Left MFG, Left AntTemp, Left PostTemp, and Left AngG (similar to Fedorenko et al, 2010; in this case, 6 parcels in left hemisphere were created based on a group-level probabilistic activation overlap map for a sentences>nonwords contrast in 220 adult participants; parcels downloaded from https://evlab.mit.edu/funcloc/). We also looked within the mirror of these search spaces in the right hemisphere (i.e., right hemisphere language homologues). We used the same method as above to define ss-fROIs for theory of mind. In this case, the ToM ss-fROI definition task was the ToM Localizer (Dodell-Feder et al., 2011) using the False Belief > False Photo contrast. The predefined ToM search spaces included 7 regions based on a group random effects analysis with 462 adult participants (Dufour et al., 2013): right and left temporoparietal junction (RTPJ, LTPJ), the precuneus (PC), the dorsal, middle and ventral components of the medial prefrontal cortex (DMPFC, MMPFC and VMPFC), and the right superior temporal sulcus (RSTS). Using the ss-fROIs defined based on the localizer tasks, we then extracted the average magnitude per condition from the SS-BlockedLang task, averaged across all usable runs per participant. Statistical analyses were conducted in R. Conditions were compared using linear mixed effects models; t-tests use Satterthwaite's method. First we tested for network-level fixed effects, with ROI and participants modeled as random effects, using: lmer(mean_topvoxels_extracted~b_or_f*d_or_m+(1|ROI)+(1|participantID), REML = FALSE), where b_or_f is backwards or forwards, d_or_m is dialogue or monologue, and ROI is region of interest within the network. Significance was determined at a level of $p<.05$ Bonferroni corrected for the 3 networks tested. To test for interactions within individual regions, we used: lmer(mean_topvoxels_extracted~b_or_f*d_or_m+(1|participantID), REML = FALSE). Significance was determined at a level of $p<.05$ Bonferroni corrected for the number of

ROIs (6 for canonical language regions, 6 for right hemisphere language regions, and 7 for ToM regions).

Individual Regions of Interest Analysis for Comprehensible Conversation: Based on the group whole brain analysis, we performed exploratory analyses in conversation regions of interest, i.e. regions that responded most to comprehensible dialogue in the whole-brain interaction. To do this, we extracted significant clusters using the uncorrected $p<.001$ thresholded group whole brain contrast for [Forward Dialogue > Backward Dialogue] > [Forward Monologue > Backward Monologue], for clusters with at least 10 voxels. We created 10mm spheres around the center of gravity (COG) for each cluster. To create ss-fROIs, an in-lab script iteratively used the z-stat image of each 3/4 combined runs (i.e., each 'fold') to determine the top 100 voxels for a given subject, ROI, and contrast (in this case, the dialogue interaction contrast). We then used the cope image from the left-out run of a given iteration to extract the betas per condition from these selected top voxels. Statistical analyses were conducted in R. Conditions were compared using linear mixed effects models; t-tests use Satterthwaite's method. To test for interactions within regions, we used: lmer(mean_topvoxels_extracted~b_or_f*d_or_m+(1|participantID), REML = FALSE).

Overlap Analysis: To determine whether the SS-BlockedLang localizer recruited canonical language regions for the Forward>Backward contrast, we compared responses to the Intact>Degraded contrast from the auditory language localizer. First, across all subjects, we quantified the overlap at the whole-brain level using the group random effects analysis results. Specifically, we calculated Dice coefficients of similarity to capture the extent of overlap in thresholded activation maps, using the formula: Dice coefficient = $2*V_{overlap}/(V_1+V_2)$, where $V_{overlap}$ refers to the number of supra-

threshold voxels identified in both tasks, V_1 and V_2 refer to the number of supra-threshold voxels for each of the two tasks, respectively (Rombouts et al, 1997; Wilson et al, 2017). We used a threshold of p<.001 (TFCE corrected) and cluster threshold of k>=10 voxels. Dice coefficients can be described as: low (0-.19), low-moderate (.2-.39), moderate (.4-.59), high-moderate (.6-.79), and high (.8-1) (Wilson et al., 2017). We also calculated the overlap at the whole brain level for each individual subject using a threshold of p<.001. Just as we expected overlap across the whole brain, we also expected a high degree of overlap between the Forward>Backward and Intact>Degraded contrasts within each language region, per subject. We calculated and report Dice coefficients to capture the extent of overlap within each language search space, using a threshold of p<.001 and cluster threshold of k>= 10 voxels to identify suprathreshold voxels. Finally, we calculated the overlap in the top 100 language-selective voxels (e.g., our definition of an ss-fROI) in each subject using the LIT task and using the auditory language localizer task, as a measure of overlap in how these tasks would define language ss-fROIs.

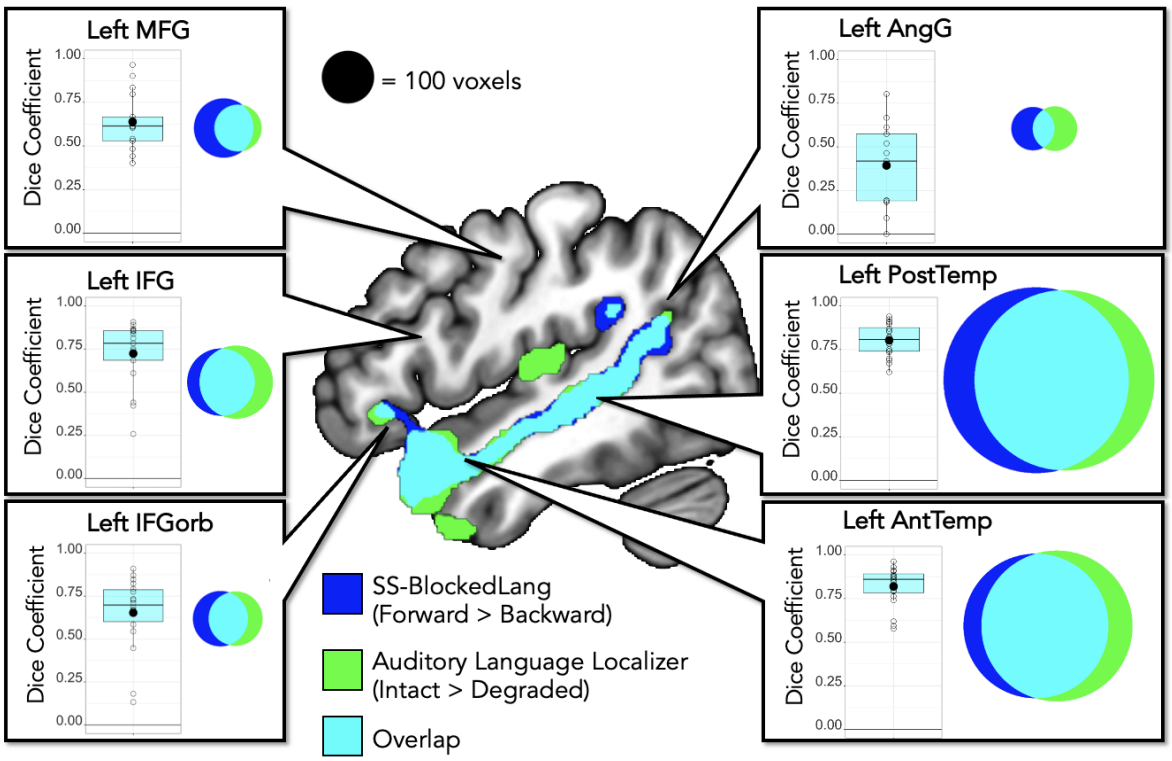
Results

Overlap between tasks

The contrast of Forward versus Backward speech in the *Sesame Street* clips robustly activated the same cortical regions as the contrast of Intact versus Degraded speech using the classic auditory language localizer, both at the group level (Dice coefficient=.71; high-moderate overlap) and at the level of individual subjects, across the whole brain and within language parcels (**Figure 2.2**; **Supplementary Tables 1-2, 6**). The mean overlap for all participants was classified as high in left AntTemp and

PostTemp regions, high-moderate in left IFGorb, IFG, and MFG, and low-moderate in left AngG.

Figure 2.2: Spatial overlap between SS-BlockedLang and Auditory Language Localizer for language contrast.

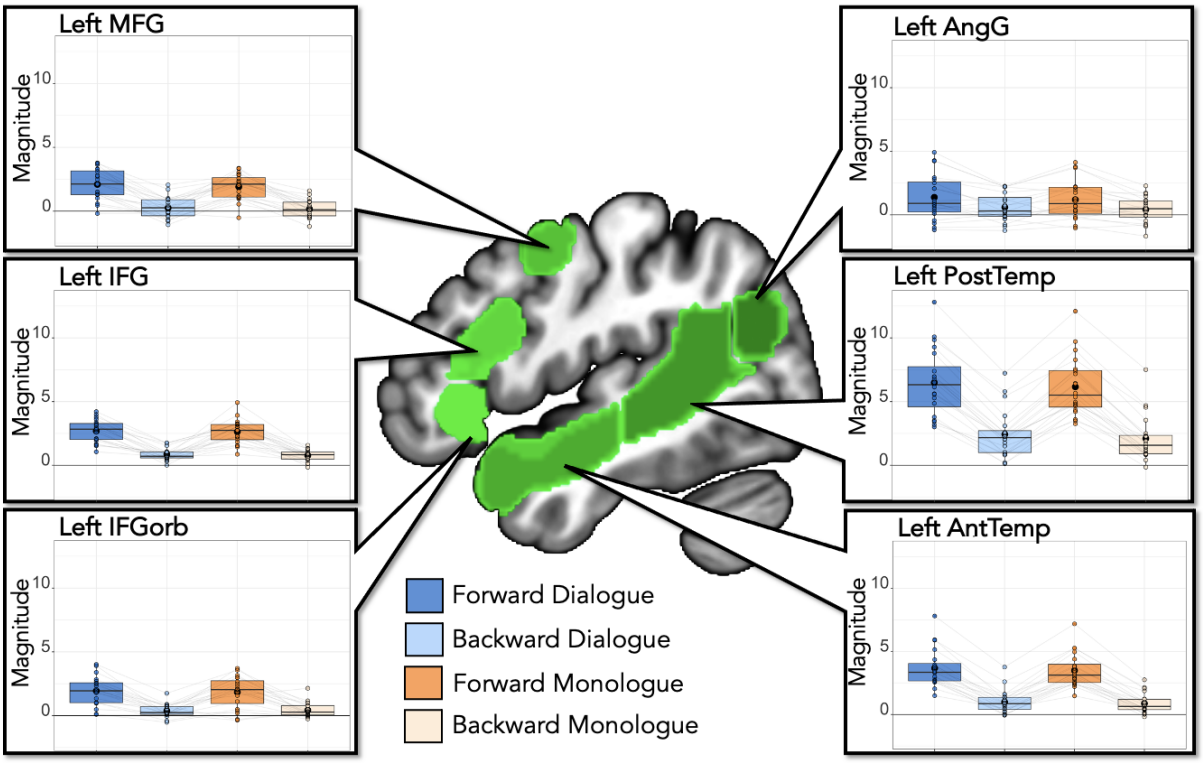


Center: *Whole brain group random effects analysis for SS-BlockedLang (Forward > Backward; blue) and the Auditory Language Localizer (Intact > Degraded; green) shows high overlap (cyan). Threshold: p<.001, TFCE corrected.* ***Panels:*** *Boxplots show the Dice coefficient (overlap) within each parcel for the language contrast using SS-BlockedLang and the Auditory Language Localizer. Threshold z>3.09; mean shown as filled black circle; open circles represent individual participants. Participants without any suprathreshold voxels in at least one contrast for a given parcel do not have a Dice coefficient and are not shown: NA = 2 for IFGorb, NA = 3 for IFG, NA = 3 for MFG, NA = 7 for AngG. Venn diagrams represent the mean number of suprathreshold voxels per contrast, per language parcel (blue: SS-BlockedLang, green: Auditory Language Localizer, cyan: overlap). Mean overlap for participants was in the high-moderate to high range for all regions except left AngG which was low-moderate overlap.*

Univariate response to task conditions in language regions

Canonical language network, including all six left-hemisphere language regions defined by the independent auditory language localizer (Scott et al., 2017), showed robust responses to both Forward speech conditions of SS-BlockedLang compared to both Backward speech conditions, as expected (Backward>Forward: Est.=-2.12, S.E.=.15, t-value=-13.69, corrected p-value<.001). This pattern held within each individual ss-fROI (**Figure 2.3**; **Table 2.1**; corrected p-values<.001 in every region). There was no main effect of Dialogue compared to Monologue in canonical language network (Dialogue>Monologue: Est.=.17, S.E.=.15, t-value=1.13, corrected p-value=.78), nor an interaction between comprehensibility and dialogue (Backward>Forward*Dialogue>Monologue: Est.=-.05, S.E.=.22, t-value=-.22, corrected p-value=1; individual language ss-fROI results in **Figure 2.3; Table 2.1**; corrected p-values>.1 in every region for dialogue and interaction).

Figure 2.3: SS-BlockedLang average magnitude by condition within language regions.



***Center**: Left hemisphere language parcels overlaid on template brain (green; parcels include left IFGorb, IFG, MFG, AntTemp, PostTemp, and AngG from https://evlab.mit.edu/funcloc/). **Panels**: Average response magnitude (betas) per individual for each condition in the SS-BlockedLang task was extracted from subject-specific functional regions of interest for language (blue: Forward Dialogue; light blue: Backward Dialogue; orange: Forward Monologue; light orange: Backward Monologue). Boxplot with mean in black circle; colored circles show individual participants with light gray lines connecting single participants. There was a main effect of Forward speech compared to Backward speech, but no effect of Dialogue speech compared to Monologue speech within language regions.*

Table 2.1: SS-BlockedLang statistics in language regions.

ROI	Backward v. Forward	Dialogue v. Monologue	Interaction
Left IFGorb	Est.= -1.57	Est.= 0.10	Est.= -0.15

	S.E.= 0.20 t-value= -7.69 p< .001 *	S.E.= 0.20 t-value=0.49 p= 0.63	S.E.= 0.29 t-value= -0.51 p= 0.61
Left IFG	Est.= -1.82 S.E.= 0.16 t-value= -11.23 p< .001 *	Est.= 0.06 S.E.= 0.16 t-value= 0.35 p= 0.73	Est.= 0.04 S.E.= 0.23 t-value= 0.19 p=0.85
Left MFG	Est.= -1.84 S.E.= 0.19 t-value= -9.77 p< .001 *	Est.=0.19 S.E.= 0.19 t-value= 1.00 p= 0.32	Est.= -0.10 S.E.= 0.27 t-value= -0.39 p= 0.70
Left AntTemp	Est.= -2.68 S.E.= 0.16 t-value= -16.26 p< .001 *	Est.= 0.19 S.E.= 0.16 t-value= 1.13 p= 0.26	Est.= -0.07 S.E.= 0.23 t-value= -0.29 p= 0.78
Left PostTemp	Est.= -4.06 S.E.= 0.21 t-value= -19.12 p< .001 *	Est.= 0.33 S.E.= 0.21 t-value= 1.56 p= 0.12	Est.= -0.01 S.E.= 0.30 t-value= -0.03 p= 0.98
Left AngG	Est.= -0.79 S.E.= 0.19 t-value= -4.25 p< .001 *	Est.= 0.19 S.E.= 0.19 t-value= 1.01 p= 0.32	Est.= -0.01 S.E.= 0.26 t-value= -0.05 p= 0.96

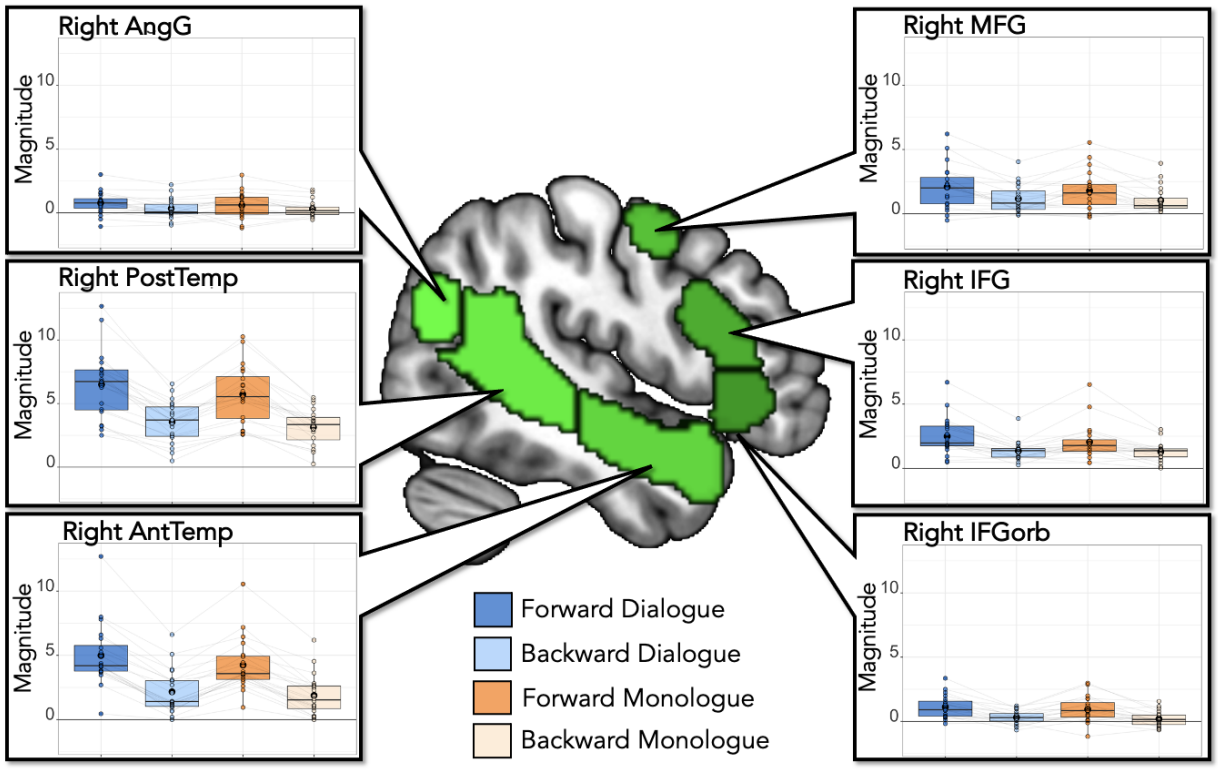
Within each language ss-fROI, there was a significant difference between Forward and Backward speech, but no difference between Monologue and Dialogue, and no interaction. Results (Est. = estimate, S.E. = standard error, t-value, and uncorrected p-value) from the model:
*lmer(mean_topvoxels_extracted~b_or_f*d_or_m+(1|participantID), REML = FALSE)*
** indicates significance level p<.05, Bonferroni corrected for 6 ROIs (p<.0083)*

Univariate response to task conditions outside canonical language regions

There were effects of dialogue in regions of cortex outside the canonical left-hemisphere language network. First, we examined right hemisphere homologues of

language regions, which responded more to forward than backward speech (Backward>Forward: Est.=-1.51, S.E.=.17, t-value=-9.05, corrected p-value<.001), and more to dialogue than monologue speech (Dialogue>Monologue: Est.=.44, S.E.=.17, t-value=2.61, corrected p-value=.028), though showed no interaction between comprehensibility and dialogue (Backward>Forward*Dialogue>Monologue: Est.=-.26, S.E.=.24, t-value=-1.11, corrected p-value=.80). Individually, all of these regions responded more to forward than backward speech, and AntTemp and PostTemp responded more to dialogue than monologue (**Figure 2.4; Table 2.2**); there were no significant interactions between comprehensibility (forward/backwards) and dialogue (dialogue/monologue) in any individual regions.

Figure 2.4: SS-BlockedLang average magnitude by condition within right homologue language regions.



***Center**: Right hemisphere language parcels (mirror of left hemisphere parcels) overlaid on template brain (green; parcels include right IFGorb, IFG, MFG, AntTemp, PostTemp, and AngG from https://evlab.mit.edu/funcloc/). **Panels**: Average response magnitude (betas) per individual for each condition in the SS-BlockedLang task was extracted from subject-specific functional regions of interest for right language homologues (blue: Forward Dialogue; light blue: Backward Dialogue; orange: Forward Monologue; light orange: Backward Monologue). Boxplot with mean in black circle; colored circles show individual participants with light gray lines connecting single participants. There was a main effect of Forward speech compared to Backward speech in all regions, and a main effect of Dialogue speech compared to Monologue speech in right AntTemp and PostTemp.*

Table 2.2: SS-BlockedLang statistics in right hemisphere language region homologues.

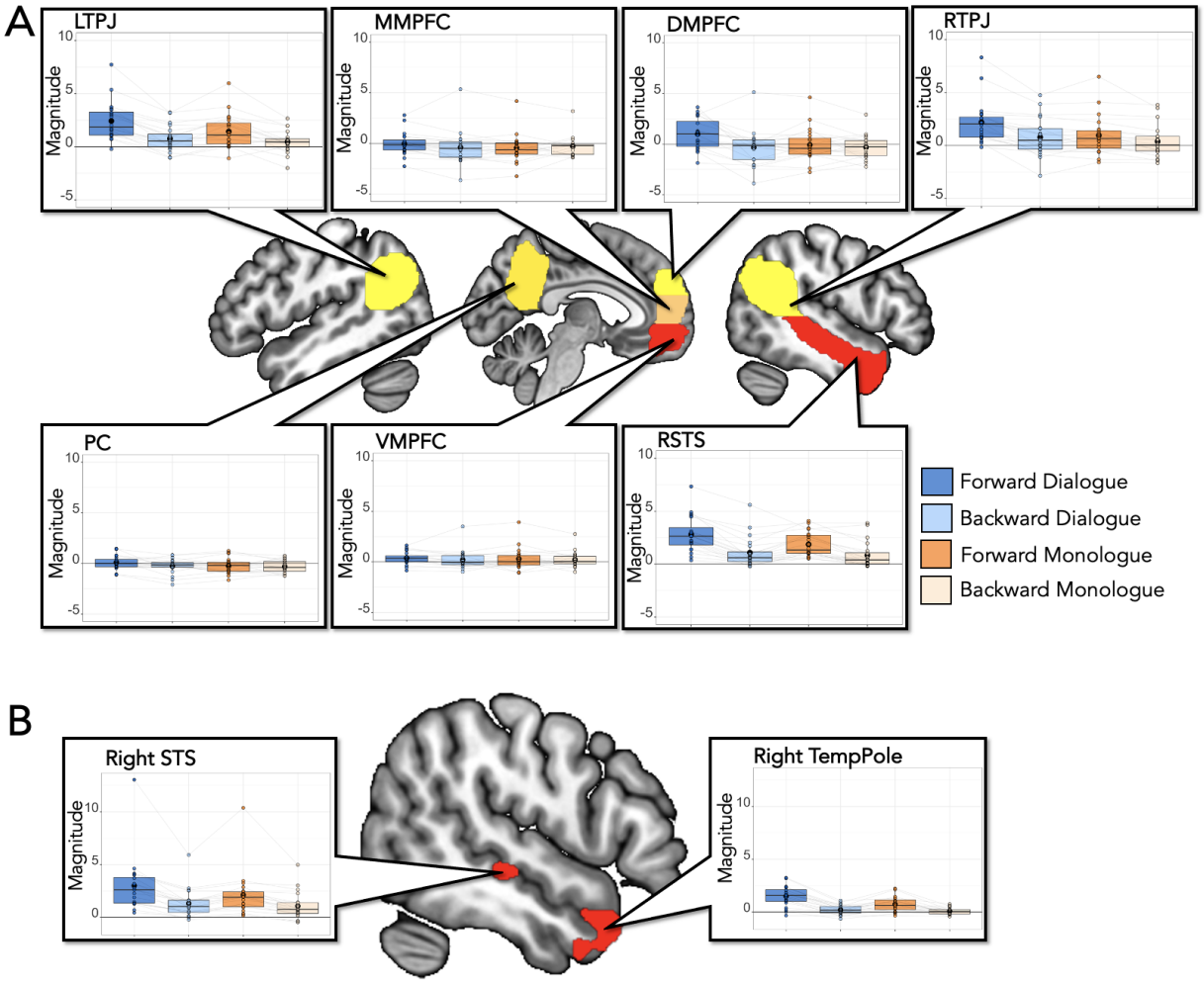
ROI	Backward v. Forward	Dialogue v. Monologue	Interaction
Right IFGorb	Est.= -0.81 S.E.= 0.15 t-value= -5.44 p< .001 *	Est.= 0.19 S.E.= 0.15 t-value= 1.26 p= 0.21	Est.= -0.07 S.E.= 0.21 t-value= -0.32 p= 0.75
Right IFG	Est.= -1.14 S.E.= 0.18 t-value= -6.46 p< .001 *	Est.= 0.44 S.E.= 0.18 t-value= 2.49 p= 0.02	Est.= -0.38 S.E.= 0.25 t-value= -1.51 p= 0.14
Right MFG	Est.= -0.92 S.E.= 0.19 t-value= -4.81 p< .001 *	Est.= 0.29 S.E.= 0.19 t-value= 1.53 p= 0.13	Est.= -0.17 S.E.= 0.27 t-value= -0.62 p= 0.54
Right AntTemp	Est.= -2.83 S.E.= 0.20 t-value= -14.31 p< .001 *	Est.= 0.74 S.E.= 0.20 t-value= 3.76 p< .001 *	Est.= -0.46 S.E.= 0.28 t-value= -1.63 p= 0.11
Right PostTemp	Est.= -2.95 S.E.= 0.26	Est.= 0.77 S.E.= 0.26	Est.= -0.36 S.E.= 0.36

	t-value= -11.55 p< .001 *	t-value= 3.03 p= 0.004 *	t-value= -1.00 p= 0.32
Right AngG	Est.= -0.43 S.E.= 0.12 t-value= -3.62 p< .001 *	Est.= 0.18 S.E.= 0.12 t-value= 1.49 p= 0.14	Est.= -0.14 S.E.= 0.17 t-value= -0.84 p= 0.40

*There was a significant difference between Forward and Backward speech within each right language homologue ss-fROI, and a main effect of Dialogue speech compared to Monologue speech in AntTemp and PostTemp, but no interaction. Results (Est. = estimate, S.E. = standard error, t-value, and uncorrected p-value) from the model: lmer(mean_topvoxels_extracted~b_or_f*d_or_m+(1|participantID), REML = FALSE)*
** indicates significance level p<.05, Bonferroni corrected for 6 ROIs (p<.0083)*

Next, we examined responses to each task condition in ToM regions. ToM network responded more to forward than backward speech (Backward>Forward: Est.=-1.00, S.E.=.15, t-value=-6.80, corrected p-value<.001), more to dialogue than monologue (Dialogue>Monologue: Est.=.73, S.E.=.15, t-value=4.93, corrected p-value<.001), and showed an interaction between comprehensibility and dialogue (Backward>Forward*Dialogue>Monologue: Est.=-.61, S.E.=.21, t-value=-2.90, corrected p-value=.012). Individually, four out of 7 regions responded more to forward than backward speech, and more to dialogue than monologue (DMPFC, LTPJ, RTPJ, and RSTS; Figure 2.5a; Table 2.3). DMPFC and RTPJ had a significant interaction between comprehensibility and dialogue, responding most for Forward Dialogue.

Figure 2.5: SS-BlockedLang whole brain interaction for comprehensible dialogue.



***(A) Center**: Theory of mind parcels overlaid on template brain (parcels include LTPJ, MMPFC, DMPFC, RTPJ, PC, VMPFC, and RSTS from (Dufour et al., 2013)). **Panels**: Average response magnitude per individual for each condition in the SS-BlockedLang task was extracted from subject-specific functional regions of interest for theory of mind (blue: Forward Dialogue; light blue: Backward Dialogue; orange: Forward Monologue; light orange: Backward Monologue). Boxplot with mean in black circle; colored circles show individual participants with light gray lines connecting single participants. There was a main effect of Forward compared to Backward speech and a main effect of Dialogue compared to Monologue in DMPFC, LTPJ, RTPJ, and RSTS, and an interaction in DMPFC and RTPJ.*

***(B) Center**: Right superior temporal sulcus and right temporal pole were activated for [Forward Dialogue > Forward Monologue] > [Backward Dialogue > Backward Monologue]. Threshold p<.001, uncorrected (df=19, two-tailed). Nothing survives at TFCE corrected threshold. Shown here are clusters for Right STS and Right Temporal*

Pole; not shown: clusters in Left STS and Left Cerebellum. ***Panels****: Average response magnitude per individual for each condition in the SS-BlockedLang task was extracted from subject-specific functional regions of interest for conversation, based on spheres around center of gravity voxels from the group whole-brain interaction contrast (blue: Forward Dialogue; light blue: Backward Dialogue; orange: Forward Monologue; light orange: Backward Monologue). Boxplot with mean in black circle; colored circles show individual participants with light gray lines connecting single participants. Results shown for Right STS and Right Temporal Pole. There was a main effect of Forward>Backward and a main effect of Dialogue>Monologue in these regions, as well as an interaction in Right Temporal Pole.*

Table 2.3: SS-BlockedLang statistics in theory of mind regions.

ROI	Backward v. Forward	Dialogue v. Monologue	Interaction
DMPFC	Est.= -1.32 S.E.= 0.27 t-value= -4.86 p< .001 *	Est.= 1.07 S.E.= 0.27 t-value= 3.92 p< .001 *	Est.= -1.08 S.E.= 0.38 t-value= -2.81 p= 0.007 *
MMPFC	Est.= -0.36 S.E.= 0.20 t-value= -1.84 p= 0.07	Est.= 0.42 S.E.= 0.20 t-value= 2.11 p= 0.04	Est.= -0.52 S.E.= 0.28 t-value= -1.87 p= 0.07
VMPFC	Est.= -0.17 S.E.= 0.14 t-value= -1.20 p= 0.24	Est.= 0.08 S.E.= 0.14 t-value= 0.58 p= 0.56	Est.= -0.10 S.E.= 0.20 t-value= -0.49 p= 0.62
LTPJ	Est.= -1.66 S.E.= 0.22 t-value= -7.73 p< .001 *	Est.= 0.99 S.E.= 0.22 t-value= 4.61 p< .001 *	Est.= -0.69 S.E.= 0.30 t-value= -2.27 p= 0.03
PC	Est.= -0.37 S.E.= 0.14 t-value= -2.74	Est.= 0.35 S.E.= 0.14 t-value= 2.56	Est.= -0.30 S.E.= 0.19 t-value= -1.59

	p= 0.008	p= 0.01	p= 0.12
RTPJ	Est.= -1.42 S.E.= 0.19 t-value= -7.37 p< .001 *	Est.= 1.25 S.E.= 0.19 t-value= 6.47 p< .001 *	Est.= -0.85 S.E.= 0.27 t-value= -3.12 p= 0.003 *
RSTS	Est.= -1.72 S.E.= 0.19 t-value= -9.21 p< .001 *	Est.= 0.94 S.E.= 0.19 t-value= 5.06 p< .001 *	Est.= -0.69 S.E.= 0.26 t-value= -2.62 p= 0.01

*Within ToM ss-fROIs, there was a main effect of Forward compared to Backward speech and a main effect of Dialogue compared to Monologue in DMPFC, LTPJ, RTPJ, and RSTS, and an interaction in DMPFC and RTPJ. Results (Est. = estimate, S.E. = standard error, t-value, and uncorrected p-value) from the model: lmer(mean_topvoxels_extracted~b_or_f*d_or_m+(1|participantID), REML = FALSE)*
** indicates significance level p<.05, Bonferroni corrected for 7 ROIs (p<.0071)*

Finally, to empirically test for regions that specifically respond to comprehensible dialogue, we performed a whole brain analysis for the following interaction: [Forward Dialogue>Forward Monologue]>[Backward Dialogue>Backward Monologue] (**Figure 2.5b**). Four clusters were identified using an uncorrected threshold of p<.001 in the right temporal pole, right STS, left STS, and left cerebellum (none survived TFCE correction for multiple comparisons). In exploratory analyses, we extracted activity in individual participants in individually defined ss-fROIs (within the 10mm sphere search spaces around center of gravity coordinates from the group results), using a leave-one-run-out approach (**Figure 2.5b; Table 2.4**). All four regions responded more to Dialogue than Monologue, and all except left Cerebellum responded more to Forward than Backward speech. There was an interaction between comprehensibility and dialogue in the right Temporal Pole and left Cerebellum.

Table 2.4: SS-BlockedLang comprehensible dialogue regions.

ROI	Voxels	MAX T-value	Peak X,Y,Z (mm)	COG X,Y,Z (mm)	Backward v. Forward	Dialogue v. Monologue	Interaction
Right Temporal Pole	299	6.64	54, 18, -26	52, 12.6, -30.9	Est.= -0.65 S.E.= 0.13 t-value= -4.95	Est.= 0.77 S.E.= 0.13 t-value= 5.82	Est.= -0.64 S.E.= 0.19 t-value= -3.41
Right STS	98	6.33	50, -26, -6	52, -23.9, -6.18	Est.= -1.06 S.E.= 0.25 t-value= -4.27	Est.= 0.87 S.E.= 0.25 t-value= 3.53	Est.= -0.62 S.E.= 0.35 t-value= -1.79
Left Crus 2 (Cerebellum)	29	4.61	-26, -78, -34	-28.1, -80, -34.6	Est.= -0.18 S.E.= 0.13 t-value= -1.43	Est.= 0.46 S.E.= 0.13 t-value= 3.59	Est.= -0.48 S.E.= 0.18 t-value= -2.65
Left STS	14	4.74	-50, -30, -4	-50.3, -30.6, -3.42	Est.= -1.20 S.E.= 0.17 t-value= -7.08	Est.= 0.54 S.E.= 0.17 t-value= 3.18	Est.= -0.46 S.E.= 0.24 t-value= -1.92

*Significant clusters at p<.001 (uncorrected, df=19, two-tailed). Peak coordinates and center of gravity (COG) for the cluster (weighted average of the coordinates by the intensities within the cluster). No significant voxels at p<.001 TFCE corrected. Within ss-fROIs defined based on 10mm spheres around the group cluster COG coordinates, there was a higher response to Forward than Backward speech in all regions except Left Cerebellum, and a main effect of Dialogue speech compared to Monologue speech in all regions, as well as an interaction in Right Temporal Pole and Left Cerebellum. Results (Est. = estimate, S.E. = standard error, t-value, and uncorrected p-value) from the model: lmer(mean_topvoxels_extracted~b_or_f*d_or_m+(1|participantID), REML = FALSE). P-values not reported since analyses were exploratory.*

Summary

These results suggest that dialogue does not modulate language processing in canonical left hemisphere cortical language regions. The magnitude of response in classic language regions appears to be determined by the presence of local structure in linguistic stimuli (common to both Forward conditions), while distinct cortical regions

are sensitive to the differences between dialogue and monologue speech, including some theory of mind regions (DMPFC, LTPJ, RTPJ, and RSTS) and right language homologues (right AntTemp and PostTemp), as well as other regions identified by exploratory whole-brain analyses (in right Temporal Pole, right STS, left Crus II in cerebellum, and left STS).

Next, we decided to further probe the sensitivity of language regions to features of dialogue by using longer clips of dialogue with interleaved forward and backward speech in Experiment 2. Rather than blocks of all-forward and all-backward speech, one character's audio stream was played forward, while the other character's audio stream was played backward (which character was forward versus backward was counterbalanced between participants). This approach complements Experiment 1 in a few ways. First, we measured canonical language regions' responses to forward speech within the temporal structure of natural dialogue, i.e. frequent short utterances, instead of long blocks. Second, the longer dialogue clips in Experiment 2 allowed us to use ISC analyses to directly measure the influence of linguistic structure, compared to all other visual and abstract semantic structure of the dialogue, on the timecourse of activity in canonical language regions.

Experiment 2: SS-IntDialog

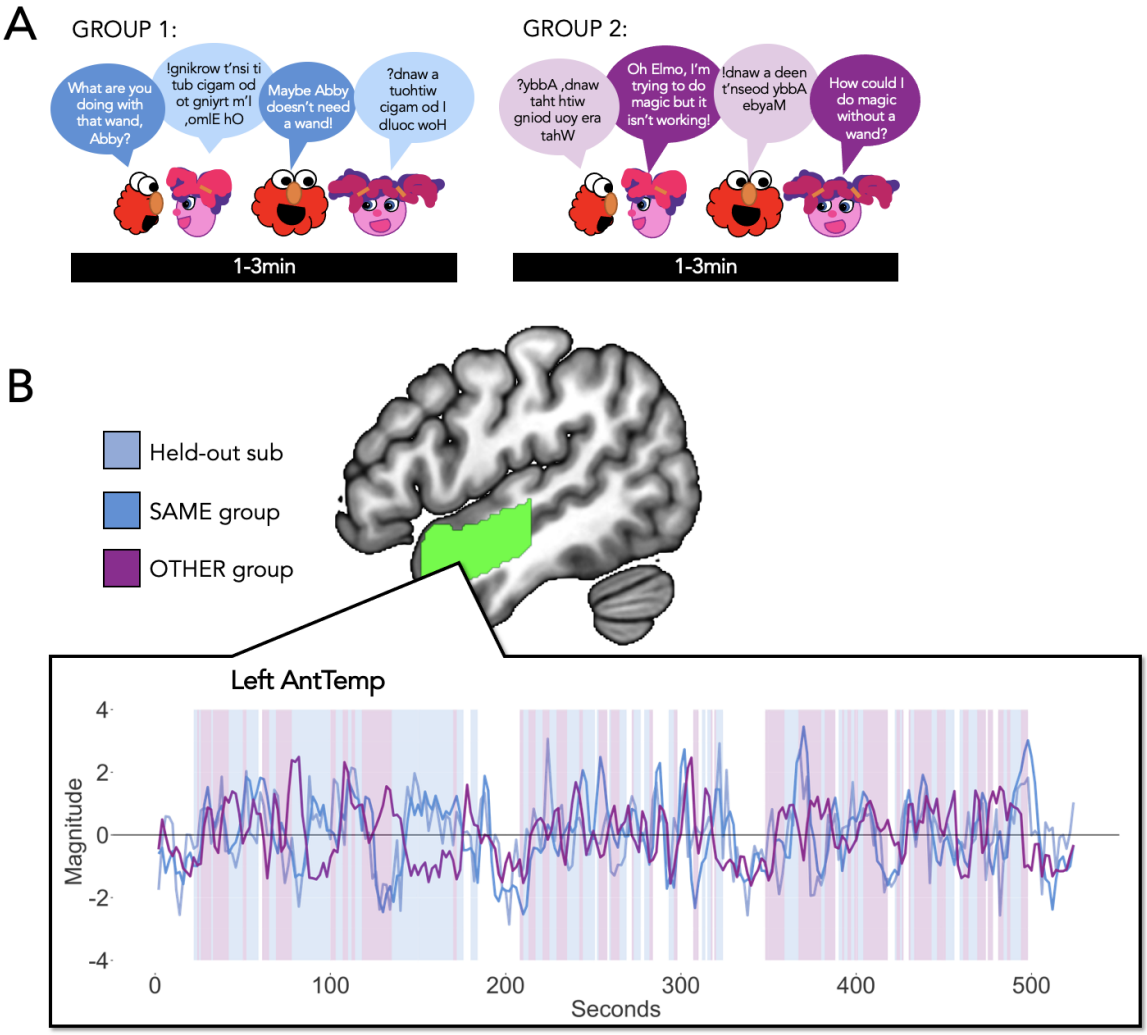
In Experiment 2, we ran our novel localizer with 1-3 minute segments of dialogue, with forward and backward speech alternating by character (SS-IntDialog), to measure responses to dialogue over time in subject-specific functional regions of interest for language and ToM.

Methods

Participants, experimental protocol, and fMRI tasks, and fMRI processing are identical to Experiment 1. In the same session, we also had participants undergo a second novel fMRI task (SS-IntDialog) that used clips of longer conversational interactions.

Stimuli Design: General methods for stimuli design were similar to Experiment 1. Our goal was to create a set of stimuli that were more naturalistic than the first task, but still allowed us to introduce an experimental manipulation of forward compared to backward speech. To do this, we selected full scenes of dialogue from *Sesame Street* during which two puppets speak to each other (the selected scenes ranged from 1-3 minutes, and we played the entire scene). Like the clips used in the SS-BlockedLang task, these scenes varied in terms of their visual properties (such as objects and setting), topic, and characters. For each clip, we reversed the audio for one character's utterances, but left the other character's audio forward (Figure 2.6A). We had two versions of each clip, such that one group of participants heard one character forward (e.g., Elmo forward and Abby backward) and the other group of participants heard the other character forward (e.g., Abby forward and Elmo backward). This allowed us to calculate ISCs between a held-out subject's timecourse and (1) the average timecourse for other participants who heard the same version of the videos, and (2) the average timecourse for the participants who heard the opposite version of the videos, within ss-fROIs (Figure 2.6B). Comprehensible utterances varied in length from .46 to 34.68 seconds, with a mean(SD) of 3.74(3.84) seconds (Figure 2.6B).

Figure 2.6: SS-IntDialog Task Design.



(A) Participants watched 1-3 minute clips of Sesame Street in which two characters have a conversation. The audio from one character is played Forward while the second is played Backward. Participants were randomly assigned to hear one of the two versions (with opposite characters played Forward/Backward). Participants watched two runs, each containing 3 clips with 20 seconds of fixation before and after each clip.
***(B) Center**: One language ROI (Left AntTemp, green). ss-fROIs were created per subject within language parcels, theory of mind parcels, and conversation spherical parcels. **Within box, left**: Example timecourse for one run of SS-IntDialog, for one participant (light blue), the average of the other participants who heard the exact same version of the run (darker blue), and the average of the participants who heard the opposite version of the run (purple). Background shading indicates when speech is*

forward (blue) or backward (purple) from the perspective of the held-out participant (opposite for the "other" group: purple is forward and blue is backward).

SS-IntDialog Language Task

Participants watched 1–3-minute dialogue clips of *Sesame Street* in which one character's audio stream was played Forward and the other was played Backward. Additional sounds in the video (e.g., blowing bubbles, a crash from something falling) were played forwards. Participants watched the videos and pressed a button on an in-scanner button box when they saw a still image of Elmo appear on the screen immediately after each block. Participants completed 2 runs, each approximately 8 min 52 sec long. Each run contained unique clips, and participants never saw a version of the same clip with the Forward/Backward streams reversed. Each run contained 3 clips, 1-3 minutes each, presented in the same order. Between each video, as well as at the beginning and end of the run, there was a 22-second fixation block. Versions of each clip with the opposite character Forward and Backward were counterbalanced between participants (randomly assigned Set A or Set B). 11 participants saw version A, and 9 participants saw version B (1 run from group A was excluded due to participant falling asleep, and one run from group B was excluded due to motion). Run order was randomized for each participant (random sequence 1-2). Transcripts and stimuli features can be found on OSF[12].

Intersubject Correlation Analysis: For the SS-IntDialog task, each participant saw two runs, each of which contained 3 different video clips (in the same order within a run). Half the participants saw version A, and half of the participants saw version B of these

[12] https://osf.io/whsb7/

runs (same videos, different audio streams). That is, if Elmo is speaking forward in the first clip in Run 1 version A, Elmo would be speaking backward in the first clip in Run 1 version B. We performed ISC analyses across the entire run, including the rest blocks between clips. ISC analyses were performed using in-lab scripts modeled after the tutorials in https://naturalistic-data.org/ (Chang et al., 2020). The preprocessed data was smoothed with a 6mm kernel, and then denoised using a GLM (6 realignment parameters, their squares, their derivatives, and squared derivatives), with outliers excluded using a dummy code, and average CSF activity and linear and quadratic trends regressed out. The timecourse was z-transformed to be centered at 0.

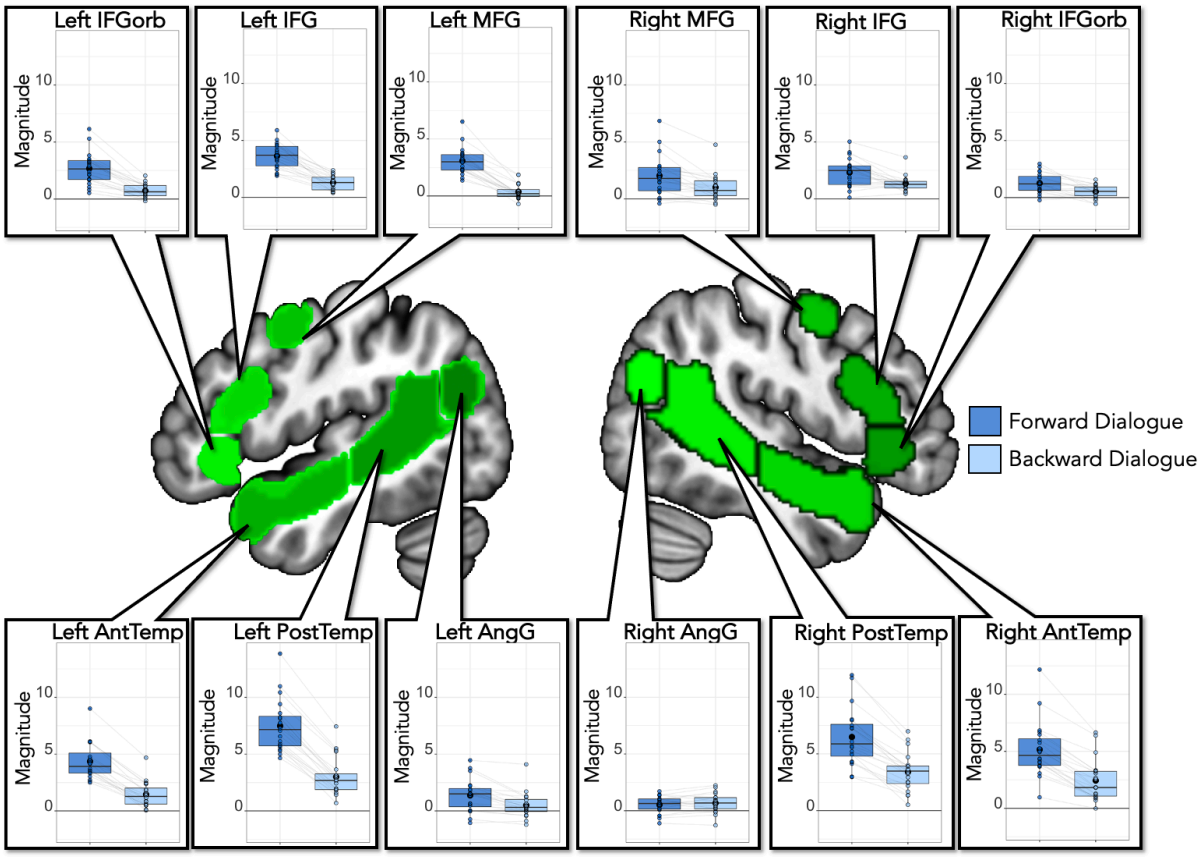
First, we extracted the timecourse per participant, per run for each language ss-fROI (defined as specified in **Exp. 1**, using the auditory language localizer). Using a leave-one-subject out approach, we calculated the correlation between the held-out subject's timecourse (i.e. the average response of that subject across all 100 voxels in that ROI) and (1) the average timecourse of the remaining participants who watched the same version of the stimuli, and (2) the average timecourse of the participants who watched the opposite version of the stimuli, for each language region. Next, we did the same analyses using the extracted timecourses per participant, per run for each of the ToM ss-fROIs. Finally, we repeated the same analysis with the extracted timecourses per participant, per run for each conversation ss-fROI, defined as the top 100 voxels for the [Forward Dialogue>Forward Monologue]>[Backward Dialogue>Backward Monologue] interaction contrast within 10mm spheres centered at the center of gravity point for each significant cluster in the group map (**Table 2.4**).

Results

Univariate response to forward and backward speech

By modeling the onset and offset of each utterance within the extended SS-IntDialog dialogues, we replicated the robust response to Forward utterances, and the very low response to Backward utterances, in canonical left-hemisphere language network (Backward>Forward: Est.= -2.54, S.E.= 0.16, t-value=-15.53, corrected p-value<.001), and in individual language regions (**Figure 2.7**, **Supplementary Table 7**). Right hemisphere homologues of language regions likewise responded more to Forward than Backward speech at a network level (Backward>Forward: Est.= -1.40, S.E.=.19, t-value= -7.46, corrected p-value<.001), and at the level of individual regions with the exception of right AngG (**Figure 2.7**, **Supplementary Table 7**). Both when looking across the whole brain and looking within language parcels, there was moderate to high-moderate overlap between the Forward>Backward contrast from SS-IntDialog and the Intact>Degraded contrast from the auditory language localizer (**Supplementary Tables 1, 4**). Thus, even with the linguistic content of the other speaker's utterance removed, canonical language regions still responded more to comprehensible than incomprehensible speech.

Figure 2.7: SS-IntDialog average magnitude by condition within language regions and right language regions homologues.



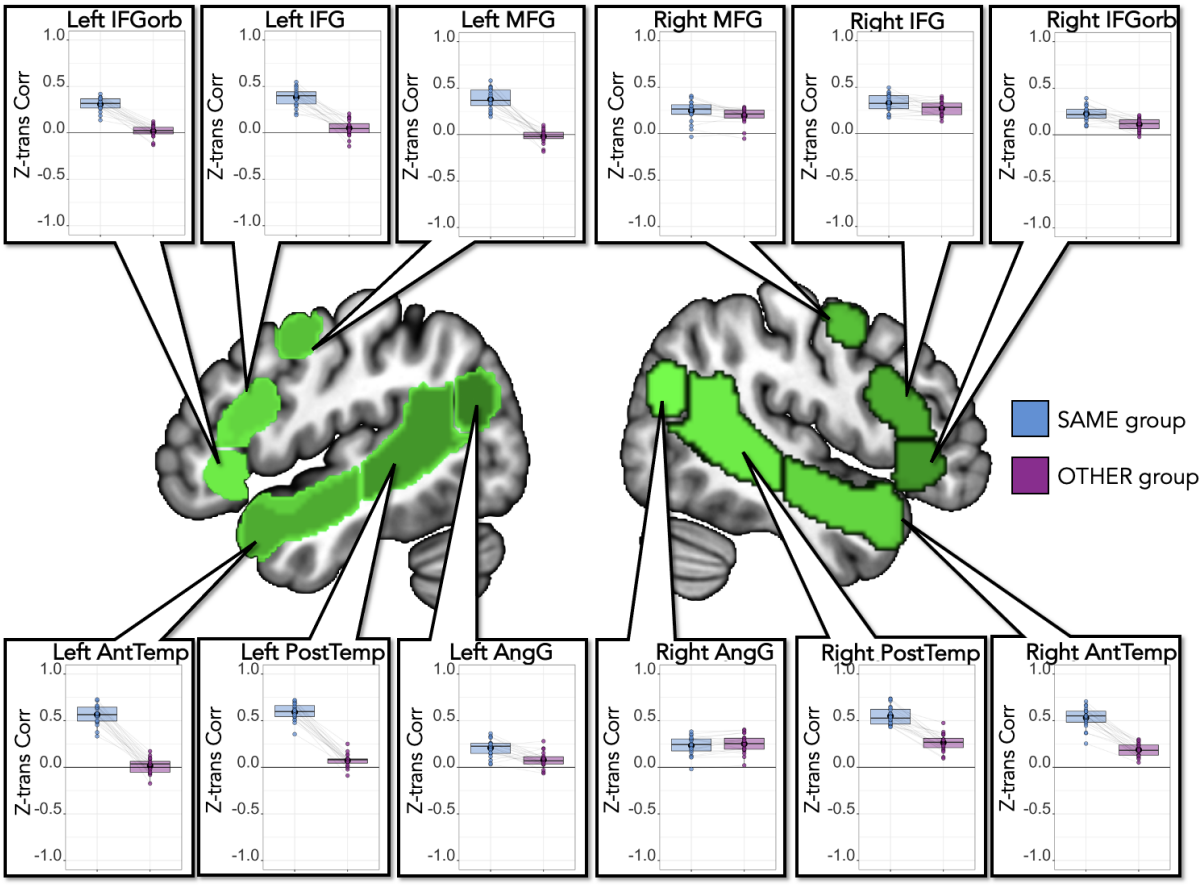
***Center**: Left hemisphere language parcels and right-hemisphere homologues overlaid on template brain (green; parcels include left and right IFGorb, IFG, MFG, AntTemp, PostTemp, and AngG from https://evlab.mit.edu/funcloc/). **Panels**: Average response magnitude per individual for each condition in the SS-IntDialog task was extracted from subject-specific functional regions of interest for language (blue: Forward Dialogue; light blue: Backward Dialogue). All regions except right AngG responded more to Forward than Backward speech.*

Timecourse of response to dialogue videos

In addition to looking at the magnitude of the response to Forward and Backward speech during dialogue, we also compared the timecourse of the response across

participants. The timecourse of response in canonical left hemisphere language regions was correlated across participants who saw the same version of the extended dialogue, with the same character's speech played forward. Thus, even the short and variable utterances within these dialogues could evoke reliable responses, consistently across participants. When comparing the timecourse to participants hearing the opposite character's speech played forward, there was little to no correlation in language regions (Figure 2.8; Table 2.5). This suggests that responses were driven by language comprehensibility, rather than visual and abstract semantic structure of the dialogues preserved between the groups (e.g., the sequence of visual images, the topic of the conversation, etc.).

Figure 2.8: SS-IntDialog correlations within language regions and right homologues.



***Center**: Left hemisphere language parcels and right-hemisphere homologues overlaid on template brain (green; parcels include left and right IFGorb, IFG, MFG, AntTemp, PostTemp, and AngG from https://evlab.mit.edu/funcloc/). **Panels**: Average z-transformed Pearson's correlation between each held-out subject's timecourse within each ss-fROI and the average timecourse of the remaining participants who viewed and listened to the same version of the stimuli (blue) and the average of the participants who heard the opposite audio stream (purple). Each individual's datapoints are connected by light gray lines. Within-group correlations were higher than between-group correlations in all regions except right IFG and AngG.*

Table 2.5: SS-IntDialog timecourse correlations within language regions.

ROI	Within-Group Correlation	Between-Group Correlation	Paired T-test

Left IFGorb	M(SD) = 0.31(0.07); range = 0.14-0.42 One-sample t-test: t-value = 19.02, **p< .001 ***	M(SD) = 0.02(0.06); range = -0.13-0.12 One-sample t-test: t-value = 1.19, p-value = 0.25	t = 14.32, **p< .001 ***
Left IFG	M(SD) = 0.38(0.10); range = 0.19-0.55 One-sample t-test: t-value = 17.30, **p< .001 ***	M(SD) = .05(0.09); range = -.14-.21 One-sample t-test: t-value = 2.59, p-value = 0.02	t = 10.02, **p< .001 ***
Left MFG	M(SD) = 0.38(0.11); range = 0.19-0.58 One-sample t-test: t-value = 15.07, **p< .001 ***	M(SD) = -0.02(0.07); range = -0.18-0.10 One-sample t-test: t-value = -1.27, p-value = 0.22	t = 11.95, **p< .001 ***
Left AntTemp	M(SD) = 0.57(0.11); range = 0.33-0.73 One-sample t-test: t-value = 24.01, **p< .001 ***	M(SD) = 0.02(0.08); range = -0.17-0.17 One-sample t-test: t-value = 0.81, p-value = 0.43	t = 18.32, **p< .001 ***
Left PostTemp	M(SD) = 0.59(0.09); range = 0.36-0.72 One-sample t-test: t-value = 28.96, **p< .001 ***	M(SD) = 0.07(0.07); range = -0.09-0.25 One-sample t-test: t-value = 4.53, **p< .001 ***	t = 22.36, **p< .001 ***
Left AngG	M(SD) = 0.21(0.09); range = 0.03-0.36 One-sample t-test:	M(SD) = 0.08(0.08); range = -0.06-0.28 One-sample t-test:	t = 4.73, p< .001 *

	t-value = 9.77, **p< .001 ***	t-value = 4.20, **p< .001 ***	

Average z-transformed Pearson's correlations between each held-out subject and the average of the rest of the group that heard the same version of the clips (within-group) and the average of the group that heard the opposite version of the clips. One-sample t-test shows significance test for two-tailed t-test against 0 (uncorrected p-values reported). Paired t-test shows that there were higher within-group than between-group correlations for each canonical language region (uncorrected p-values reported).
** indicates significance level p<.05, Bonferroni corrected for 6 ROIs (p<.0083)*

Table 2.6: SS-IntDialog timecourse correlations within right language region homologues.

ROI	Within-Group Correlation	Between-Group Correlation	Paired T-test
Right IFGorb	M(SD) = 0.23(0.08); range = 0.09-0.39 One-sample t-test: t-value = 12.44, **p< .001 ***	M(SD) = 0.11(0.07); range = -0.02-0.21 One-sample t-test: t-value = 7.46, **p< .001 ***	t = 5.36, p< .001 *
Right IFG	M(SD) = 0.33(0.10); range = 0.17-0.50 One-sample t-test: t-value = 15.39, **p< .001 ***	M(SD) = 0.27(0.08); range = 0.13-0.40 One-sample t-test: t-value = 15.64, **p< .001 ***	t = 2.89, p-value = 0.009
Right MFG	M(SD) = 0.24(0.12); range = -0.03-0.41 One-sample t-test: t-value = 9.41, **p< .001 ***	M(SD) = 0.19(0.09); range = -0.05-0.29 One-sample t-test: t-value = 9.92, **p< .001 ***	t = 3.00, p-value = 0.007 *
Right AntTemp	M(SD) = 0.530.11); range = 0.25-0.71	M(SD) = 0.19(0.07); range = 0.05-0.30	t = 14.01, **p< .001 ***

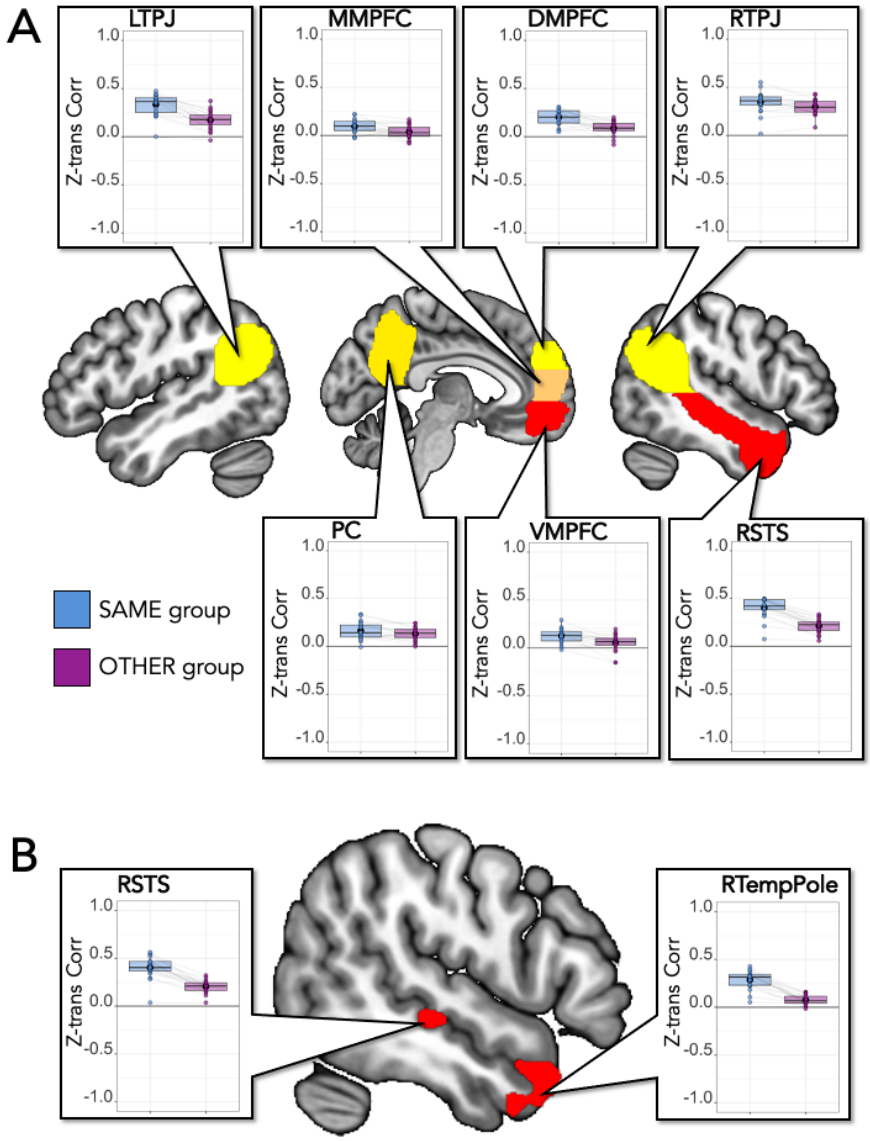
	One-sample t-test: t-value = 21.39, p< .001 *	One-sample t-test: t-value = 11.71, p< .001 *	
Right PostTemp	M(SD) = 0.55(0.10); range = 0.43-0.74 One-sample t-test: t-value = 24.37, p< .001 *	M(SD) = 0.26(0.09); range = 0.10-0.48 One-sample t-test: t-value = 12.67, p< .001 *	t = 10.62, p< .001 *
Right AngG	M(SD) = 0.23(0.10); range = -0.02-0.38 One-sample t-test: t-value = 10.25, p< .001 *	M(SD) = 0.25(0.10); range = 0.02-0.40 One-sample t-test: t-value = 11.46, p< .001 *	t = -1.39, p-value = 0.18

Average z-transformed Pearson's correlations between each held-out subject and the average of the rest of the group that heard the same version of the clips (within-group) and the average of the group that heard the opposite version of the clips. One-sample t-test shows significance test for two-tailed t-test against 0 (uncorrected p-values reported). Paired t-test shows that there were higher within-group than between-group correlations for each right hemisphere language region except right IFG and AngG (uncorrected p-values reported).
** indicates significance level p<.05, Bonferroni corrected for 6 ROIs (p<.0083)*

The dialogue videos also evoked shared temporal structure across and within participant groups in other cortical regions. In right hemisphere language regions, both the within-group and between-group Pearson's correlations were greater than 0 (**Figure 2.8; Table 2.6**), though the within-group correlations were higher than the between-group correlations for all of the regions except right IFG and AngG. Brain regions defined by the independent ToM localizer also showed significant correlations in the timecourse of response with the same and also with opposite videos (with the

exception of VMPFC and MMPFC), again with higher within-group correlations than between-group correlations for all regions except MMPFC and PC (Figure 2.9A; Table 2.7). The brain regions in right STS, right temporal pole, left STS, and left cerebellum identified as specifically responsive to comprehensible dialogue in SS-BlockedLang similarly showed correlated timecourses when participants watched matched or opposite videos (with the exception of left STS for the between-groups correlation; Figure 2.9B; Table 2.8), but had higher correlations for the matched videos. Thus, the preserved visual and abstract semantic structure of the dialogue contributed to reliable cortical responses outside of the canonical left hemisphere language regions.

Figure 2.9: SS-IntDialog correlations within theory of mind regions.



***(A) Center**: Theory of mind parcels overlaid on template brain (parcels include LTPJ, MMPFC, DMPFC, RTPJ, PC, VMPFC, and RSTS from (Dufour et al., 2013)). **Panels**: Average z-transformed Pearson's correlation between each held-out subject's timecourse within each ss-fROI and the average timecourse of the remaining participants who viewed and listened to the same version of the stimuli (blue) and the average of the participants who heard the opposite audio stream (purple), averaged across two runs. Each individual's datapoints are connected by light gray lines. For all regions except MMPFC and PC, the within-group correlations were higher than the between-group correlations.*

(B) Center: Right superior temporal sulcus and right temporal pole were activated for [Forward Dialogue > Forward Monologue] > [Backward Dialogue > Backward Monologue]. Threshold p<.001, uncorrected (df=19, two-tailed). Nothing survives at TFCE corrected threshold. Shown here are clusters for Right STS and Right Temporal Pole; not shown: Left STS and Left Cerebellum. Panels: Average z-transformed Pearson's correlation between each held-out subject's timecourse within each ss-fROI and the average timecourse of the remaining participants who viewed and listened to the same version of the stimuli (blue) and the average of the participants who heard the opposite audio stream (purple), averaged across two runs. Each individual's datapoints are connected by light gray lines.

Table 2.7: SS-IntDialog timecourse correlations within theory of mind regions.

ROI	Within-Group Correlation	Between-Group Correlation	Paired T-test
DMPFC	M(SD) = 0.20(0.08) ; range = 0.05-0.31 One-sample t-test: t-value = 10.90, p< .001 *	M(SD) = 0.09(0.07); range = -0.08-0.20 One-sample t-test: t-value = 5.39, p< .001 *	t = 4.54, p< .001*
MMPFC	M(SD) = 0.09(0.07); range = -0.03-0.22 One-sample t-test: t-value = 5.66, p< .001*	M(SD) = 0.036(0.07); range = -0.08-0.16 One-sample t-test: t-value = 2.28, p-value = 0.03	t = 2.83, p-value = 0.01
VMPFC	M(SD) = 0.13(0.08); range = -0.02-0.29 One-sample t-test: t-value = 7.47, p< .001*	M(SD) = 0.05(0.09); range = -0.15-0.20 One-sample t-test: t-value = 2.64, p-value = 0.02	t = 3.66, p-value = 0.002*
LTPJ	M(SD) = 0.33(0.11); range = -0.0007-0.47	M(SD) = 0.17(0.10); range = -0.04-0.37	t = 8.13, p< .001 ***

	One-sample t-test: t-value = 13.04, p< .001 *	One-sample t-test: t-value = 7.91, p< .001 *	
PC	M(SD) = 0.17(0.09); range = -0.005-0.34 One-sample t-test: t-value = 8.09, p< .001 *	M(SD) = 0.13(0.07); range = 0.001-0.24 One-sample t-test: t-value = 8.38, p< .001 *	t = 2.00, p-value = 0.06
RTPJ	M(SD) = 0.34(0.11); range = 0.02-0.55 One-sample t-test: t-value = 13.54, p< .001 *	M(SD) = 0.30(0.08); range = 0.08-0.43 One-sample t-test: t-value = 16.26, p< .001 *	t = 3.02, p-value = 0.007 *
RSTS	M(SD) = 0.40(0.11); range = 0.08-0.50 One-sample t-test: t-value = 16.78, p< .001 *	M(SD) = 0.21(0.07); range = 0.06-0.34 One-sample t-test: t-value = 12.90, p< .001*	t = 11.57, p< .001 *

Average z-transformed Pearson's correlations between each held-out subject and the average of the rest of the group that heard the same version of the clips (within-group) and the average of the group that heard the opposite version of the clips. One-sample t-test shows significance test for two-tailed t-test against 0 (uncorrected p-values). Paired t-test shows that there were higher within-group than between-group correlations for each ToM region except MMPFC and PC (uncorrected p-values).
** indicates significance level p<.05, Bonferroni corrected for 7 ROIs (p<.0071)*

Table 2.8: SS-IntDialog timecourse correlations within conversation regions.

ROI	Within-Group Correlation	Between-Group Correlation	Paired T-test
RTempPole	M(SD) = 0.29(0.10); range = 0.05-0.43 One-sample t-test: t-value = 13.34	M(SD) = 0.08(0.05); range = -0.01-0.16 One-sample t-test: t-value = 6.79	t = 10.28

RSTS	M(SD) = 0.40(0.12); range = 0.04-0.56 One-sample t-test: t-value = 15.45	M(SD) = 0.21(0.06); range = 0.04-0.32 One-sample t-test: t-value = 14.21	t = 10.01
LCere	M(SD) = 0.26(0.08); range = 0.12-0.38 One-sample t-test: t-value = 14.56	M(SD) = 0.17(0.08); range = 0.003-0.28 One-sample t-test: t-value = 9.69	t = 4.0738
LSTS	M(SD) = 0.40(0.14); range = 0.01-0.57 One-sample t-test: t-value = 12.95	M(SD) = 0.03(0.07); range = -0.10-0.12 One-sample t-test: t-value = 1.87	t = 10.16

Average z-transformed correlations between each held-out subject and the average of the rest of the group that heard the same version of the clips (within-group) and the average of the group that heard the opposite version of the clips. One-sample t-test shows significance test for two-tailed t-test against 0 (uncorrected). Paired t-test shows higher within-group than between-group correlations within each region. No p-values are reported since analyses were exploratory.

Summary

Even when some of the auditory content of dialogue was removed by alternating comprehensible and incomprehensible speech across interleaved speakers, canonical left-hemisphere cortical language regions still responded more to forward than backward speech. Furthermore, the timecourses of activity in canonical left-hemisphere language regions were more similar among individuals listening to the same speech input, compared to individuals listening to the opposite auditory input, holding constant all visual input from the dialogue videos. Notably, there was little to no similarity in the response of canonical language regions for individuals listening to

opposite auditory streams, suggesting that language regions were insensitive to any other aspects of the dialogue stimuli that were the same across participants. On the other hand, multiple theory of mind regions and right hemisphere homologues of language regions were correlated even when participants were listening to the opposite auditory streams, suggesting that these regions tracked similarities in the dialogue videos in addition to just the comprehensibility of the speech (though all except right IFG, right AngG, MMPFC, and PC still had higher correlations when participants were listening to the same speech stream).

General Discussion

In this study, we tested the scope and limits of language regions' function by probing them with naturalistic videos of dialogue speech. Canonical language network is highly specific to language processing (Braga et al., 2020; Fedorenko et al., 2010, 2011; Ivanova et al., 2020; Liu et al., 2020; Pritchett et al., 2018). Dialogue, compared to monologue, is a useful boundary test case: tracking multiple speakers and perspectives is part of comprehending language in dialogue, but it is not a function that has been attributed to language regions.

In two tasks, we manipulated the audio stream of *Sesame Street* videos to create matched segments of videos with forward (comprehensible) and backward (incomprehensible) speech. Based on this manipulation, we defined three measures of a cortical region's (in)sensitivity to the dialogue context of linguistic input. First, a region that processes language independent of a dialogue context should respond equally robustly to forward speech whether presented as a monologue or dialogue. Second, it should respond selectively to the comprehensible speech segments in a

dialogue that alternates between forward and backwards speech, even within the frequent alternations of dialogue that render some utterances quite short. Third, the reliable (between-participants) timecourse of response to these alternating dialogue stimuli should be driven only by the timing of the comprehensible speech segments, and not by any other features of the dialogue. By all three of these measures, we find that left hemisphere canonical language regions are insensitive to whether language is in the form of dialogue during passive observation.

Insensitivity to dialogue in canonical language regions

We chose two different analytic approaches to ascertain whether language regions are sensitive to the back-and-forth speaker alternation in dialogue. Using a blocked design, we found no differences in the magnitude of neural response to dialogue versus monologue in the core left-hemisphere language network, though other cortical regions did show enhanced responses during dialogue (see below). With the longer audiovisual stimuli in the SS-IntDialog task, we employed a different analytic technique (ISC) that is more commonly used for naturalistic fMRI datasets (Hasson et al., 2004). ISC analyses allowed us to ask: does anything other than language comprehensibility, at the level of individual utterances, drive the neural responses in language regions during a dialogue? Participants who watched the exact same video clips had similar responses to the stimuli in language regions (i.e., positive within-group correlations within language regions). Critically, though, participants who watched the same video clips with the opposite characters speaking forward versus backward showed close to zero correlation in canonical language network activity. Thus, nothing else about the dialogue, other than the comprehensibility of the speech stream, was reliably tracked by canonical language regions across participants. This did not have to be the case, as we saw in other regions (discussed below).

Our two analytic approaches, and two task designs, both produced results that converged on a single conclusion: canonical language regions are not sensitive features of dialogue other than language. This insensitivity is consistent with other evidence that language regions are sensitive to relatively local linguistic features, and with evidence that canonical language regions have fairly short temporal receptive windows (Blank & Fedorenko, 2020).

Sensitivity to dialogue outside language regions

In addition to canonical language regions, we also asked whether other regions might be specifically sensitive to dialogue. First, we looked in specific regions that one might expect to respond differently to dialogue and monologue: ToM regions and right hemisphere homologues of language regions. Some regions in both groups fulfilled our criteria for responding preferentially to dialogue.

Most individually-defined ToM regions responded more during forward than backward speech, and more during two-character dialogue interactions than a single character speaking. Two regions responded most during comprehensible dialogue, as shown by an interaction of language and video type. As converging evidence, in SS-IntDialog, activity in ToM regions was also correlated across participants even when they were listening to the opposite speech streams in Experiment 2 (with the exception of MMPFC and VMPFC), which suggests that aspects of the stimuli other than comprehensibility influence activity in these regions. In experiments using single source texts, ToM regions respond selectively to stimuli that describe or imply contrasting beliefs, knowledge or emotions, between characters or over time (Dodell-Feder et al., 2011; Saxe & Kanwisher, 2003; Saxe & Powell, 2006). Naturalistic dialogue often

implies differences of perspective, as speakers use utterances to show how a prior utterance was or was not understood, and to reveal and correct gaps in common ground. Previous work has shown that activity in ToM regions can be synchronized with language regions during language comprehension (Paunov et al., 2019) even though these networks are functionally distinct (Paunov et al., 2022; Shain et al., 2022); thus, it is unsurprising that ToM network was engaged in processing dialogue in our task.

The other set of regions we examined was right language homologues. We localized these regions in the same way as the left hemisphere regions - using the independent language localizer task, selecting top voxels from parcels in the right hemisphere. All of these regions did respond more to comprehensible than incomprehensible speech, and 2 of the 6 regions also responded more to dialogue than monologue. More intriguingly, all of the right hemisphere language homologues showed significant correlations in SS-IntDialog even between participants listening to opposite versions of the stimuli. Thus, dialogue videos evoke enhanced activity in some right hemisphere, but not left hemisphere, regions defined by their language selectivity. Consistent with these results, previous work has demonstrated that pragmatic and social aspects of language may be processed by regions in the right hemisphere. For instance, processing emotional prosody has been shown to be right lateralized (Friederici, 2011; Frühholz et al., 2012; Ross & Monnot, 2008; Seydell-Greenwald et al., 2020), and regions responsive to prosody differences have been shown to be distinct from language regions, even among individuals with large perinatal strokes in the left hemisphere whose language regions are located in the right hemisphere (Newport et al., 2022). Right hemisphere damage can make it more difficult for individuals to make inferences from discourse (Beeman, 1993).

Finally, in addition to looking within specific regions, we also examined responses across the whole brain to determine where comprehensible dialogue specifically led to higher activation. Significant clusters were identified in right temporal pole, right STS, left STS, and left cerebellum (though note that none of these survived correction for multiple comparisons). While these results are exploratory, these regions may be useful targets for future studies on dialogue comprehension. Part of right temporal pole, for example, has long been thought to be involved in social and emotional processing, among other higher level cognitive functions (Herlin et al., 2021; Olson et al., 2007; Pehrs et al., 2017; Wakusawa et al., 2007), and parts of the cerebellum are also involved in language and social cognition (D'Mello & Stoodley, 2015; Stoodley, 2012; Van Overwalle et al., 2014); thus, it is plausible that the clusters we identified in these regions might be meaningful subregions involved in dialogue comprehension. Right STS in particular is a key region supporting social interaction processing, though it is important to note that this is a large region with multiple subregions subserving different functions (Deen et al., 2015). Parts of STS respond to visual social interactions (Walbrin et al., 2018; Walbrin & Koldewyn, 2019), and in particular, a specific part of posterior STS responds to interactions between agents (Isik et al., 2017). Using a naturalistic dataset, part of STS was shown to be selective for interaction, separate from TOM (Lee Masson & Isik, 2021). Other evidence points to additional roles of STS regions in social processing, such as directing attention (Materna et al., 2008) and processing prosody (Wildgruber et al., 2006). Given the interactive nature of the dialogue stimuli, it makes sense that part of STS might be involved in processing the social interaction that occurs in a comprehensible dialogue, either as a subregion that responds to both social interaction and voices, or because the content of the language enhances the perception of an interaction.

Limitations

In these experiments, the stimuli were experimentally manipulated clips from professionally produced episodes of the television show, *Sesame Street*. This stimulus source had both strengths and limitations, in terms of experimental design for testing the function of language regions. First, there were some differences in the linguistic features between the dialogue and monologue clips (**Supplementary Figure 4**). For example, monologues included longer sentences and more complex speech than dialogue. Second, we used puppets with rigid mouths, to allow us to align backwards speech with the video stimuli. Both children and adults are highly sensitive to cross modal misalignment of lip movements and speech sounds. However, there may be residual differences between conditions in the audiovisual alignment, since the puppets were originally filmed to match the forward speech stream. Third, we used backwards speech as the control condition rather than acoustically degraded speech (Overath et al., 2015; Stoppelman et al., 2013) or foreign speech (Schlosser et al., 1998). We chose to use backwards speech because it (1) allowed us to maintain the continuity of voice within character, (2) it sounded more natural than the degraded speech when embedded in the video, and (3) most critically, because it allowed us to control for the visual stimuli across participants (e.g., we could have different participants watch the exact same clips, either forward or backward), which would not have been feasible given the availability of non-English versions of the selected clips. Future studies could design stimuli with comprehensible and foreign speech using bilingual actors to match the visual input and voices.

Future Directions

A clear extension of this work - and indeed, the motivation behind it - is to use these language tasks with young children. Extensive prior literature has demonstrated the

benefits of naturalistic movie-based stimuli for young children (Cantlon, 2020; Cantlon & Li, 2013; Kamps et al., 2022; Redcay & Moraczewski, 2020; Richardson et al., 2018; Vanderwal et al., 2015, 2019). It was non-trivial that this quasi-naturalistic approach worked at all for univariate analyses: one concern we had was that the uncontrolled elements of the video clips would prevent us from extracting a robust language response. This was not the case: not only did our block design in Experiment 1 elicit a response with substantial overlap with a more traditional language localizer - both at the group and individual level - but even the naturalistic alternating dialogue in Experiment 2 reliably localized language regions. In each participant, the same voxels were identified by the forward vs backward manipulation of the *Sesame Street* clips as by a well-validated standard auditory language localizer task (Scott et al., 2017). Furthermore, this child-friendly task was engaging and effective for adults, suggesting that it holds promise as a task that can be used across a wide age range. Thus, this task may be useful for other populations that may find classic language tasks hard to tolerate, such as individuals with developmental or acquired disorders.

Conclusions

Our results suggest that canonical left hemisphere language regions are not sensitive to aspects of dialogue interactions other than comprehensibility of the speech stream, even though other aspects of dialogue can impact language comprehension. Furthermore, we found that embedding an explicit experimental control for language (in our case, backwards speech) within a naturalistic context (in our case, clips from *Sesame Street*) is a feasible, engaging, and robust approach for studying language processing in the brain.

Materials

Stimuli transcriptions and descriptions, analysis code, stimulus presentation code, processed data, and link to raw data on OpenNeuro can be found on OSF (https://osf.io/whsb7/). Raw stimuli can be provided upon request.

Acknowledgements

We would like to thank Somaia Saba, Hana Ro, and Michelle Hung for their assistance with stimuli creation. Thank you to Ev Fedorenko, Shari Liu, and Nancy Kanwisher for helpful feedback on this manuscript. Thank you to Steve Shannon and Atsushi Takahashi at the Athinoula A. Martinos Imaging Center at MIT. Finally, thank you to our participants for making this research possible.

Funding Sources: This research was supported by the Simons Foundation Autism Research Initiative via the Simons Center for the Social Brain at MIT and the NSF Graduate Research Fellowship Program (#1745302 to HO).

References

Bašnáková, J., Weber, K., Petersson, K. M., van Berkum, J., & Hagoort, P. (2014). Beyond the Language Given: The Neural Correlates of Inferring Speaker Meaning. *Cerebral Cortex*, *24*(10), 2572–2578. https://doi.org/10.1093/cercor/bht112

Bates, E., Reilly, J., Wulfeck, B., Dronkers, N., Opie, M., Fenson, J., Kriz, S., Jeffries, R., Miller, L., & Herbst, K. (2001). Differential Effects of Unilateral Lesions on Language Production in Children and Adults. *Brain and Language*, *79*(2), 223–265. https://doi.org/10.1006/brln.2001.2482

Beeman, M. (1993). Semantic Processing in the Right Hemisphere May Contribute to Drawing Inferences from Discourse. *Brain and Language*, *44*(1), 80–120. https://doi.org/10.1006/brln.1993.1006

Bendtz, K., Ericsson, S., Schneider, J., Borg, J., Bašnáková, J., & Uddén, J. (2022). Individual Differences in Indirect Speech Act Processing Found Outside the Language Network. *Neurobiology of Language*, *3*(2), 287–317. https://doi.org/10.1162/nol_a_00066

Binder, J. R., Frost, J. A., Hammeke, T. A., Cox, R. W., Rao, S. M., & Prieto, T. (1997). Human Brain Language Areas Identified by Functional Magnetic Resonance Imaging. *Journal of Neuroscience*, *17*(1), 353–362. https://doi.org/10.1523/JNEUROSCI.17-01-00353.1997

Blank, I. A., & Fedorenko, E. (2020). No evidence for differences among language regions in their temporal receptive windows. *NeuroImage*, *219*, 116925. https://doi.org/10.1016/j.neuroimage.2020.116925

Blank, I., Balewski, Z., Mahowald, K., & Fedorenko, E. (2016). Syntactic processing is distributed across the language system. *NeuroImage*, *127*, 307–323. https://doi.org/10.1016/j.neuroimage.2015.11.069

Blank, I., Kanwisher, N., & Fedorenko, E. (2014). A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. *Journal of Neurophysiology*, *112*(5), 1105–1118. https://doi.org/10.1152/jn.00884.2013

Blumstein, S. E., & Amso, D. (2013). Dynamic Functional Organization of Language: Insights From Functional Neuroimaging. *Perspectives on Psychological Science*, *8*(1), 44–48. https://doi.org/10.1177/1745691612469021

Braga, R. M., DiNicola, L. M., Becker, H. C., & Buckner, R. L. (2020). Situating the left-lateralized language network in the broader organization of multiple specialized large-scale distributed networks. *Journal of Neurophysiology*, *124*(5), 1415–1448. https://doi.org/10.1152/jn.00753.2019

Broca, P. (1865). Sur le siège de la faculté du langage articulé. *Bulletins et Mémoires de la Société d'Anthropologie de Paris*, *6*(1), 377–393. https://doi.org/10.3406/bmsap.1865.9495

Cantlon, J. F. (2020). The balance of rigor and reality in developmental neuroscience. *NeuroImage*, *216*, 116464. https://doi.org/10.1016/j.neuroimage.2019.116464

Cantlon, J. F., & Li, R. (2013). Neural Activity during Natural Viewing of Sesame Street Statistically Predicts Test Scores in Early Childhood. *PLOS Biology*, *11*(1), e1001462. https://doi.org/10.1371/journal.pbio.1001462

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. https://doi.org/10.1016/j.jml.2016.06.013

Chang, L., Manning, J., Baldassano, C., Vega, A. de la, Fleetwood, G., Geerligs, L., Haxby, J., Lahnakoski, J., Parkinson, C., Shappell, H., Shim, W. M., Wager, T., Yarkoni, T., Yeshurun, Y., & Finn, E. (2020, July 9). *naturalistic-data-analysis/naturalistic_data_analysis: Version 1.0*. https://doi.org/10.5281/zenodo.3937849

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

Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science*, *13*(2), 259–294. https://doi.org/10.1016/0364-0213(89)90008-6

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39. https://doi.org/10.1016/0010-0277(86)90010-7

Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015). Functional Organization of Social Perception and Cognition in the Superior Temporal Sulcus. *Cerebral Cortex*, *25*(11), 4596–4609. https://doi.org/10.1093/cercor/bhv111

D'Mello, A. M., & Stoodley, C. J. (2015). Cerebro-cerebellar circuits in autism spectrum disorder. *Frontiers in Neuroscience*, *9*. https://www.frontiersin.org/articles/10.3389/fnins.2015.00408

Dodell-Feder, D., Koster-Hale, J., Bedny, M., & Saxe, R. (2011). FMRI item analysis in a theory of mind task. *NeuroImage*, *55*(2), 705–712. https://doi.org/10.1016/j.neuroimage.2010.12.040

Dronkers, N. F., Wilkins, D. P., Van Valin, R. D., Redfern, B. B., & Jaeger, J. J. (2004). Lesion analysis of the brain areas involved in language comprehension. *Cognition*, *92*(1), 145–177. https://doi.org/10.1016/j.cognition.2003.11.002

Dufour, N., Redcay, E., Young, L., Mavros, P. L., Moran, J. M., Triantafyllou, C., Gabrieli, J. D. E., & Saxe, R. (2013). Similar Brain Activation during False Belief Tasks in a Large Sample of Adults with and without Autism. *PLoS ONE*, *8*(9), e75468. https://doi.org/10.1371/journal.pone.0075468

Enge, A., Friederici, A. D., & Skeide, M. A. (2020). A meta-analysis of fMRI studies of language comprehension in children. *NeuroImage*, *215*, 116858. https://doi.org/10.1016/j.neuroimage.2020.116858

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), Article 1. https://doi.org/10.1038/s41592-018-0235-4

Fedorenko, E., Behr, M. K., & Kanwisher, N. (2011). Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences*, *108*(39), 16428–16433. https://doi.org/10.1073/pnas.1112937108

Fedorenko, E., & Blank, I. A. (2020). Broca's Area Is Not a Natural Kind. *Trends in Cognitive Sciences*, *24*(4), 270–284. https://doi.org/10.1016/j.tics.2020.01.001

Fedorenko, E., Duncan, J., & Kanwisher, N. (2012). Language-Selective and Domain-General Regions Lie Side by Side within Broca's Area. *Current Biology*, *22*(21), 2059–2062. https://doi.org/10.1016/j.cub.2012.09.011

Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New Method for fMRI Investigations of Language: Defining ROIs Functionally in Individual Subjects. *Journal of Neurophysiology*, *104*(2), 1177–1194. https://doi.org/10.1152/jn.00032.2010

Fedorenko, E., Ivanova, A., Dhamala, R., & Bers, M. U. (2019). The Language of Programming: A Cognitive Perspective. *Trends in Cognitive Sciences*, *23*(7), 525–528. https://doi.org/10.1016/j.tics.2019.04.010

Fedorenko, E., & Thompson-Schill, S. L. (2014). Reworking the language network. *Trends in Cognitive Sciences*, *18*(3), 120–126. https://doi.org/10.1016/j.tics.2013.12.006

Feng, W., Wu, Y., Jan, C., Yu, H., Jiang, X., & Zhou, X. (2017). Effects of contextual relevance on pragmatic inference during conversation: An fMRI study. *Brain and Language*, *171*, 52–61. https://doi.org/10.1016/j.bandl.2017.04.005

Fox Tree, J. E. (1999). Listening in on monologues and dialogues. *Discourse Processes*, *27*(1), 35–53. https://doi.org/10.1080/01638539909545049

Friederici, A. D. (2011). The Brain Basis of Language Processing: From Structure to Function. *Physiological Reviews*, *91*(4), 1357–1392. https://doi.org/10.1152/physrev.00006.2011

Friederici, A. D., & Gierhan, S. M. (2013). The language network. *Current Opinion in Neurobiology*, *23*(2), 250–254. https://doi.org/10.1016/j.conb.2012.10.002

Frühholz, S., Ceravolo, L., & Grandjean, D. (2012). Specific Brain Networks during Explicit and Implicit Decoding of Emotional Prosody. *Cerebral Cortex*, *22*(5), 1107–1117. https://doi.org/10.1093/cercor/bhr184

Fusaroli, R., & Tylén, K. (2016). Investigating Conversational Dynamics: Interactive Alignment, Interpersonal Synergy, and Collective Task Performance. *Cognitive Science*, *40*(1), 145–171. https://doi.org/10.1111/cogs.12251

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, *8*(1), 8–11. https://doi.org/10.1016/j.tics.2003.10.016

Gogate, L. J., & Bahrick, L. E. (1998). Intersensory Redundancy Facilitates Learning of Arbitrary Relations between Vowel Sounds and Objects in Seven-Month-Old

Infants. *Journal of Experimental Child Psychology*, *69*(2), 133–149. https://doi.org/10.1006/jecp.1998.2438

Gold, B. T., & Buckner, R. L. (2002). Common Prefrontal Regions Coactivate with Dissociable Posterior Regions during Controlled Semantic and Phonological Tasks. *Neuron*, *35*(4), 803–812. https://doi.org/10.1016/S0896-6273(02)00800-0

Gorgolewski, K., Burns, C., Madison, C., Clark, D., Halchenko, Y., Waskom, M., & Ghosh, S. (2011). Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. *Frontiers in Neuroinformatics*, *5*. https://www.frontiersin.org/articles/10.3389/fninf.2011.00013

Grall, C., & Finn, E. S. (2022). Leveraging the power of media to drive cognition: A media-informed approach to naturalistic neuroscience. *Social Cognitive and Affective Neuroscience*, *17*(6), 598–608. https://doi.org/10.1093/scan/nsac019

Hagoort, P. (2014). Nodes and networks in the neural architecture for language: Broca's region and beyond. *Current Opinion in Neurobiology*, *28*, 136–141. https://doi.org/10.1016/j.conb.2014.07.013

Hagoort, P., & Indefrey, P. (2014). The Neurobiology of Language Beyond Single Words. *Annual Review of Neuroscience*, *37*(1), 347–362. https://doi.org/10.1146/annurev-neuro-071013-013847

Hamilton, L. S., & Huth, A. G. (2020). The revolution will not be controlled: Natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, *35*(5), 573–582. https://doi.org/10.1080/23273798.2018.1499946

Hasson, U., & Honey, C. J. (2012). Future trends in Neuroimaging: Neural processes as expressed within real-life contexts. *NeuroImage*, *62*(2), 1272–1278. https://doi.org/10.1016/j.neuroimage.2012.02.004

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject Synchronization of Cortical Activity During Natural Vision. *Science*, *303*(5664), 1634–1640. https://doi.org/10.1126/science.1089506

Herlin, B., Navarro, V., & Dupont, S. (2021). The temporal pole: From anatomy to function—A literature appraisal. *Journal of Chemical Neuroanatomy*, *113*, 101925. https://doi.org/10.1016/j.jchemneu.2021.101925

Hu, J., Small, H., Kean, H., Takahashi, A., Zekelman, L., Kleinman, D., Ryan, E., Nieto-Castañón, A., Ferreira, V., & Fedorenko, E. (2022). Precision fMRI reveals that the language-selective network supports both phrase-structure building and lexical access during language production. *Cerebral Cortex*, bhac350. https://doi.org/10.1093/cercor/bhac350

Isik, L., Koldewyn, K., Beeler, D., & Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences*, *114*(43), E9145–E9152. https://doi.org/10.1073/pnas.1714471114

Ivanova, A. A., Srikant, S., Sueoka, Y., Kean, H. H., Dhamala, R., O'Reilly, U.-M., Bers, M. U., & Fedorenko, E. (2020). Comprehension of computer code relies primarily on domain-general executive brain regions. *ELife*, *9*, e58906. https://doi.org/10.7554/eLife.58906

Jang, G., Yoon, S., Lee, S.-E., Park, H., Kim, J., Ko, J. H., & Park, H.-J. (2013). Everyday conversation requires cognitive inference: Neural bases of comprehending implicated meanings in conversations. *NeuroImage*, *81*, 61–72. https://doi.org/10.1016/j.neuroimage.2013.05.027

Kamps, F. S., Richardson, H., Murty, N. A. R., Kanwisher, N., & Saxe, R. (2022). Using child-friendly movie stimuli to study the development of face, place, and object regions from age 3 to 12 years. *Human Brain Mapping*, *43*(9), 2782–2800. https://doi.org/10.1002/hbm.25815

Lee Masson, H., & Isik, L. (2021). Functional selectivity for social interaction perception in the human superior temporal sulcus during natural viewing. *NeuroImage*, *245*, 118741. https://doi.org/10.1016/j.neuroimage.2021.118741

Levinson, S. C. (2016). Turn-taking in Human Communication – Origins and Implications for Language Processing. *Trends in Cognitive Sciences*, *20*(1), 6–14. https://doi.org/10.1016/j.tics.2015.10.010

Lewkowicz, D. J., & Flom, R. (2014). The Audiovisual Temporal Binding Window Narrows in Early Childhood. *Child Development*, *85*(2), 685–694. https://doi.org/10.1111/cdev.12142

Liu, Y.-F., Kim, J., Wilson, C., & Bedny, M. (2020). Computer code comprehension shares neural resources with formal logical inference in the fronto-parietal network. *ELife*, *9*, e59340. https://doi.org/10.7554/eLife.59340

MacSweeney, M., Capek, C. M., Campbell, R., & Woll, B. (2008). The signing brain: The neurobiology of sign language. *Trends in Cognitive Sciences*, *12*(11), 432–440. https://doi.org/10.1016/j.tics.2008.07.010

Malik-Moraleda, S., Ayyash, D., Gallée, J., Affourtit, J., Hoffmann, M., Mineroff, Z., Jouravlev, O., & Fedorenko, E. (2022). An investigation across 45 languages and 12 language families reveals a universal language network. *Nature Neuroscience*, *25*(8), Article 8. https://doi.org/10.1038/s41593-022-01114-5

Materna, S., Dicke, P. W., & Thier, P. (2008). The posterior superior temporal sulcus is involved in social communication not specific for the eyes. *Neuropsychologia*, *46*(11), 2759–2765. https://doi.org/10.1016/j.neuropsychologia.2008.05.016

Menenti, L., Gierhan, S. M. E., Segaert, K., & Hagoort, P. (2011). Shared Language: Overlap and Segregation of the Neuronal Infrastructure for Speaking and Listening Revealed by Functional MRI. *Psychological Science*, *22*(9), 1173–1182. https://doi.org/10.1177/0956797611418347

Monti, M. M., Parsons, L. M., & Osherson, D. N. (2012). Thought Beyond Language: Neural Dissociation of Algebra and Natural Language. *Psychological Science*, *23*(8), 914–922. https://doi.org/10.1177/0956797612437427

Nastase, S. A., Goldstein, A., & Hasson, U. (2020). Keep it real: Rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage*, *222*, 117254. https://doi.org/10.1016/j.neuroimage.2020.117254

Nastase, S. A., Liu, Y.-F., Hillman, H., Zadbood, A., Hasenfratz, L., Keshavarzian, N., Chen, J., Honey, C. J., Yeshurun, Y., Regev, M., Nguyen, M., Chang, C. H. C., Baldassano, C., Lositsky, O., Simony, E., Chow, M. A., Leong, Y. C., Brooks, P. P., Micciche, E., … Hasson, U. (2021). The "Narratives" fMRI dataset for evaluating models of naturalistic language comprehension. *Scientific Data*, *8*(1), Article 1. https://doi.org/10.1038/s41597-021-01033-3

Neville, H. J., Bavelier, D., Corina, D., Rauschecker, J., Karni, A., Lalwani, A., Braun, A., Clark, V., Jezzard, P., & Turner, R. (1998). Cerebral organization for language in deaf and hearing subjects: Biological constraints and effects of experience. *Proceedings of the National Academy of Sciences*, *95*(3), 922–929. https://doi.org/10.1073/pnas.95.3.922

Newport, E. L., Seydell-Greenwald, A., Landau, B., Turkeltaub, P. E., Chambers, C. E., Martin, K. C., Rennert, R., Giannetti, M., Dromerick, A. W., Ichord, R. N., Carpenter, J. L., Berl, M. M., & Gaillard, W. D. (2022). Language and developmental plasticity after perinatal stroke. *Proceedings of the National Academy of Sciences*, *119*(42), e2207293119. https://doi.org/10.1073/pnas.2207293119

Olson, I. R., Plotzker, A., & Ezzyat, Y. (2007). The Enigmatic temporal pole: A review of findings on social and emotional processing. *Brain*, *130*(7), 1718–1731. https://doi.org/10.1093/brain/awm052

Overath, T., McDermott, J. H., Zarate, J. M., & Poeppel, D. (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nature Neuroscience*, *18*(6), Article 6. https://doi.org/10.1038/nn.4021

Paunov, A. M., Blank, I. A., & Fedorenko, E. (2019). Functionally distinct language and Theory of Mind networks are synchronized at rest and during language comprehension. *Journal of Neurophysiology*, *121*(4), 1244–1265. https://doi.org/10.1152/jn.00619.2018

Paunov, A. M., Blank, I. A., Jouravlev, O., Mineroff, Z., Gallée, J., & Fedorenko, E. (2022). Differential Tracking of Linguistic vs. Mental State Content in Naturalistic Stimuli by Language and Theory of Mind (ToM) Brain Networks. *Neurobiology of Language*, 1–29. https://doi.org/10.1162/nol_a_00071

Pehrs, C., Zaki, J., Schlochtermeier, L. H., Jacobs, A. M., Kuchinke, L., & Koelsch, S. (2017). The Temporal Pole Top-Down Modulates the Ventral Visual Stream During Social Cognition. *Cerebral Cortex*, *27*(1), 777–792. https://doi.org/10.1093/cercor/bhv226

Price, C. J. (2010). The anatomy of language: A review of 100 fMRI studies published in 2009. *Annals of the New York Academy of Sciences*, *1191*(1), 62–88. https://doi.org/10.1111/j.1749-6632.2010.05444.x

Price, C. J. (2012). A review and synthesis of the first 20years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage*, *62*(2), 816–847. https://doi.org/10.1016/j.neuroimage.2012.04.062

Pritchett, B. L., Hoeflin, C., Koldewyn, K., Dechter, E., & Fedorenko, E. (2018). High-level language processing regions are not engaged in action observation or imitation. *Journal of Neurophysiology*, *120*(5), 2555–2570. https://doi.org/10.1152/jn.00222.2018

Redcay, E., & Moraczewski, D. (2020). Social cognition in context: A naturalistic imaging approach. *NeuroImage*, *216*, 116392. https://doi.org/10.1016/j.neuroimage.2019.116392

Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., & Saxe, R. (2018). Development of the social brain from age three to twelve years. *Nature Communications*, *9*(1), Article 1. https://doi.org/10.1038/s41467-018-03399-2

Ross, E. D., & Monnot, M. (2008). Neurology of affective prosody and its functional–anatomic organization in right hemisphere. *Brain and Language*, *104*(1), 51–74. https://doi.org/10.1016/j.bandl.2007.04.007

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking peopleThe role of the temporo-parietal junction in "theory of mind." *NeuroImage*, *19*(4), 1835–1842. https://doi.org/10.1016/S1053-8119(03)00230-1

Saxe, R., & Powell, L. J. (2006). It's the Thought That Counts: Specific Brain Regions for One Component of Theory of Mind. *Psychological Science*, *17*(8), 692–699. https://doi.org/10.1111/j.1467-9280.2006.01768.x

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, *53*(2), 361–382. https://doi.org/10.1353/lan.1977.0041

Schlosser, M. J., Aoyagi, N., Fulbright, R. K., Gore, J. C., & McCarthy, G. (1998). Functional MRI studies of auditory comprehension. *Human Brain Mapping*, *6*(1), 1–13. https://doi.org/10.1002/(SICI)1097-0193(1998)6:1<1::AID-HBM1>3.0.CO;2-7

Scott, T. L., Gallée, J., & Fedorenko, E. (2017). A new fun and robust version of an fMRI localizer for the frontotemporal language system. *Cognitive Neuroscience*, *8*(3), 167–176. https://doi.org/10.1080/17588928.2016.1201466

Seydell-Greenwald, A., Chambers, C. E., Ferrara, K., & Newport, E. L. (2020). What you say versus how you say it: Comparing sentence comprehension and emotional prosody processing using fMRI. *NeuroImage*, *209*, 116509. https://doi.org/10.1016/j.neuroimage.2019.116509

Shain, C., Paunov, A., Chen, X., Lipkin, B., & Fedorenko, E. (2022). *No evidence of theory of mind reasoning in the human language network* (p. 2022.07.18.500516). bioRxiv. https://doi.org/10.1101/2022.07.18.500516

Sonkusare, S., Breakspear, M., & Guo, C. (2019). Naturalistic Stimuli in Neuroscience: Critically Acclaimed. *Trends in Cognitive Sciences*, *23*(8), 699–714. https://doi.org/10.1016/j.tics.2019.05.004

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J. P., Yoon, K.-E., & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*(26), 10587–10592. https://doi.org/10.1073/pnas.0903616106

Stoodley, C. J. (2012). The Cerebellum and Cognition: Evidence from Functional Imaging Studies. *The Cerebellum*, *11*(2), 352–365. https://doi.org/10.1007/s12311-011-0260-7

Stoppelman, N., Harpaz, T., & Ben-Shachar, M. (2013). Do not throw out the baby with the bath water: Choosing an effective baseline for a functional localizer of speech processing. *Brain and Behavior*, *3*(3), 211–222. https://doi.org/10.1002/brb3.129

Thompson-Schill, S. L., D'Esposito, M., Aguirre, G. K., & Farah, M. J. (1997). Role of left inferior prefrontal cortex in retrieval of semantic knowledge: A reevaluation. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(26), 14792–14797. https://doi.org/10.1073/pnas.94.26.14792

Tolins, J., & Fox Tree, J. E. (2016). Overhearers Use Addressee Backchannels in Dialog Comprehension. *Cognitive Science*, *40*(6), 1412–1434. https://doi.org/10.1111/cogs.12278

Van Overwalle, F., Baetens, K., Mariën, P., & Vandekerckhove, M. (2014). Social cognition and the cerebellum: A meta-analysis of over 350 fMRI studies. *NeuroImage*, *86*, 554–572. https://doi.org/10.1016/j.neuroimage.2013.09.033

Vanderwal, T., Eilbott, J., & Castellanos, F. X. (2019). Movies in the magnet: Naturalistic paradigms in developmental functional neuroimaging. *Developmental Cognitive Neuroscience*, *36*, 100600. https://doi.org/10.1016/j.dcn.2018.10.004

Vanderwal, T., Kelly, C., Eilbott, J., Mayes, L. C., & Castellanos, F. X. (2015). Inscapes: A movie paradigm to improve compliance in functional magnetic resonance imaging. *NeuroImage*, *122*, 222–232. https://doi.org/10.1016/j.neuroimage.2015.07.069

Wakusawa, K., Sugiura, M., Sassa, Y., Jeong, H., Horie, K., Sato, S., Yokoyama, H., Tsuchiya, S., Inuma, K., & Kawashima, R. (2007). Comprehension of implicit meanings in social situations involving irony: A functional MRI study. *NeuroImage*, *37*(4), 1417–1426. https://doi.org/10.1016/j.neuroimage.2007.06.013

Walbrin, J., Downing, P., & Koldewyn, K. (2018). Neural responses to visually observed social interactions. *Neuropsychologia*, *112*, 31–39. https://doi.org/10.1016/j.neuropsychologia.2018.02.023

Walbrin, J., & Koldewyn, K. (2019). Dyadic interaction processing in the posterior temporal cortex. *NeuroImage*, *198*, 296–302. https://doi.org/10.1016/j.neuroimage.2019.05.027

Wehbe, L., Blank, I. A., Shain, C., Futrell, R., Levy, R., von der Malsburg, T., Smith, N., Gibson, E., & Fedorenko, E. (2021). Incremental Language Comprehension Difficulty Predicts Activity in the Language Network but Not the Multiple Demand Network. *Cerebral Cortex*, *31*(9), 4006–4023. https://doi.org/10.1093/cercor/bhab065

Wernicke, C. (1874). Der aphasische Symptomencomplex: Eine psychologische Studie auf anatomischer Basis. Cohn & Weigert.

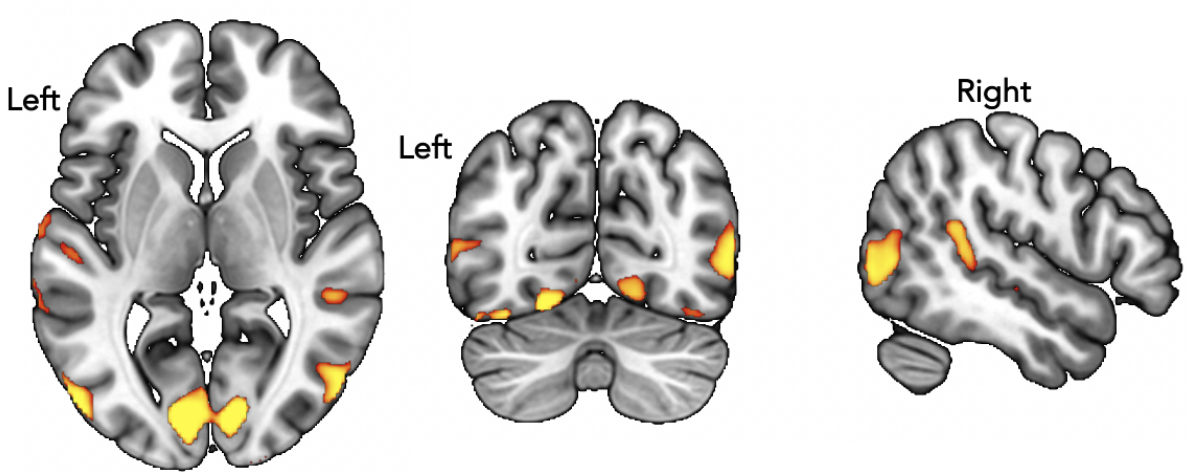
Wildgruber, D., Ackermann, H., Kreifelts, B., & Ethofer, T. (2006). Cerebral processing of linguistic and emotional prosody: FMRI studies. In S. Anders, G. Ende, M. Junghofer, J. Kissler, & D. Wildgruber (Eds.), *Progress in Brain Research* (Vol. 156, pp. 249–268). Elsevier. https://doi.org/10.1016/S0079-6123(06)56013-3

Wilson, S. M., Bautista, A., Yen, M., Lauderdale, S., & Eriksson, D. K. (2017). Validity and reliability of four language mapping paradigms. *NeuroImage: Clinical*, *16*, 399–408. https://doi.org/10.1016/j.nicl.2016.03.015

Supplementary Materials

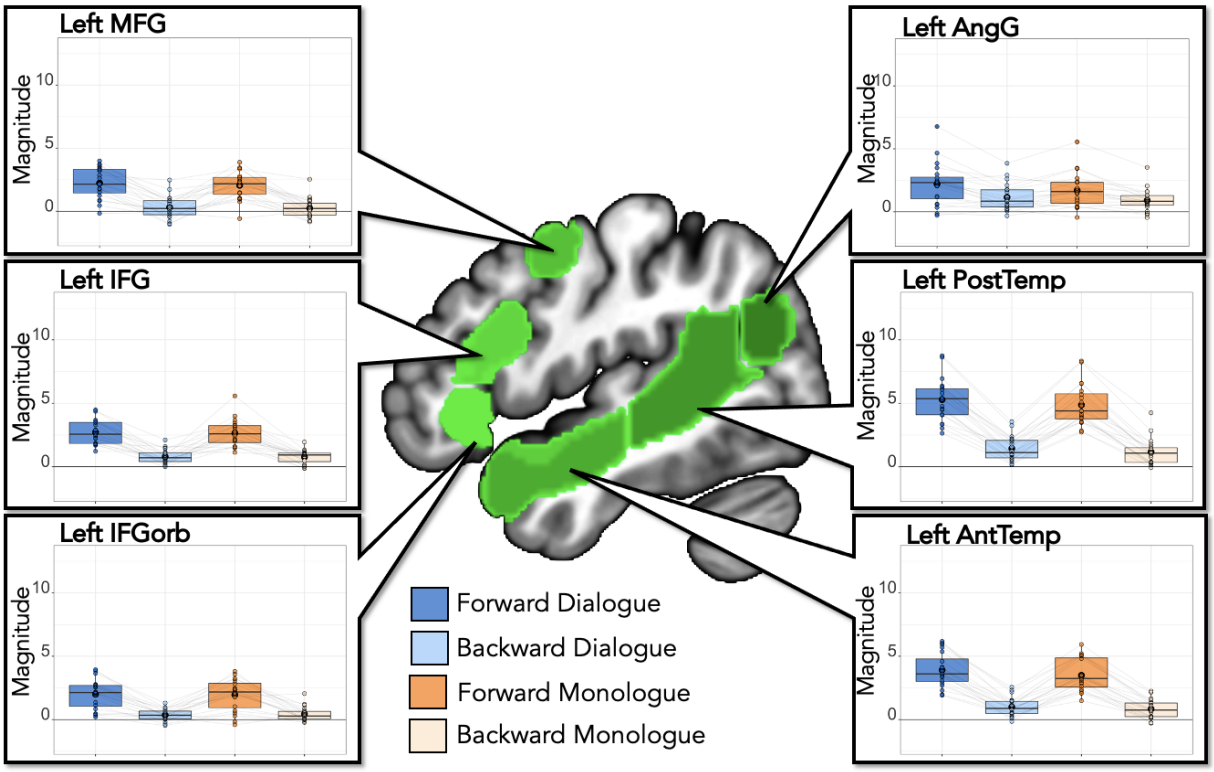
Supplemental Figures

SUPPLEMENTARY FIGURE 1: Whole brain contrast for social interaction.



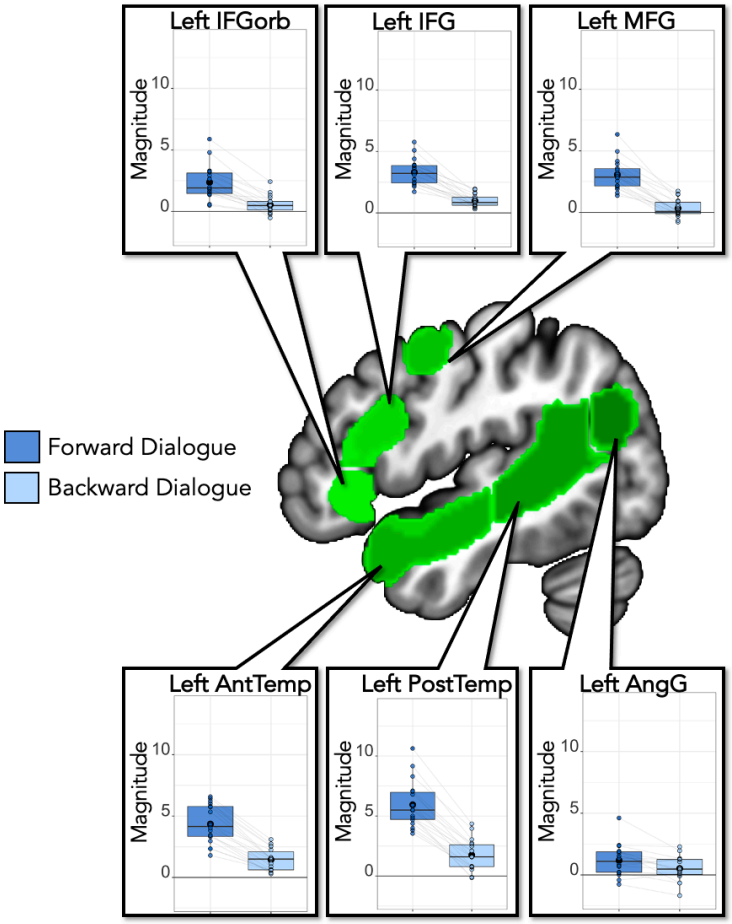
Group whole-brain analysis for the Backward Dialogue > Backward Monologue contrast in SS-BlockedLang. Shown at threshold: p<.01, TFCE corrected.

SUPPLEMENTARY FIGURE 2: SS-BlockedLang magnitude in SS-BlockedLang-defined ss-fROIs for language.



***Center**: Left hemisphere language parcels overlaid on template brain (green; parcels include left IFGorb, IFG, MFG, AntTemp, PostTemp, and AngG from https://evlab.mit.edu/funcloc/). **Panels**: Average response magnitude (betas) per individual for held-out runs in each condition in the SS-BlockedLang task was extracted from subject-specific functional regions of interest for language, defined by the SS-BlockedLang Forward>Backward contrast (blue: Forward Dialogue; light blue: Backward Dialogue; orange: Forward Monologue; light orange: Backward Monologue). Boxplot with mean in black circle; colored circles show individual participants with light gray lines connecting single participants.*

SUPPLEMENTARY FIGURE 3: SS-IntDialog magnitude in SS-IntDialog-defined ss-fROIs for language.



***Center**: Left hemisphere language parcels overlaid on template brain (green; parcels include left IFGorb, IFG, MFG, AntTemp, PostTemp, and AngG from https://evlab.mit.edu/funcloc/). **Panels**: Average response magnitude (betas) per individual for held-out runs in each condition in the SS-IntDialog task was extracted from subject-specific functional regions of interest for language, defined by the SS-IntDialog Forward>Backward contrast (blue: Forward Dialogue; light blue: Backward Dialogue). Boxplot with mean in black circle; colored circles show individual participants with light gray lines connecting single participants.*

Supplemental Tables

SUPPLEMENTARY TABLE 1: Whole brain overlap for language contrasts.

	# LIT suprathreshold voxels	# Auditory Language Localizer suprathreshold voxels	Dice Coefficient
SS-BlockedLang	M(SD) = 13466(8188.58); range = 1606-32596	M(SD) = 14403(8260.71); range = 4356-37531	M(SD) = 0.55(0.12); range = 0.17-0.76
SS-IntDialog	M(SD) = 9148(5456.05); range = 3229-20684	M(SD) = 14403(8260.71); range = 4356-37531	M(SD) = 0.50(0.09); range = 0.31-0.68

Suprathreshold voxels for each contrast, for each individual, were identified based on a threshold of z>3.09. Overlap between SS-BlockedLang and LangLoc, and between SS-IntDialog and LangLoc, across the entire brain were both in the moderate range.

SUPPLEMENTARY TABLE 2: SS-BlockedLang overlap in language parcels.

ROI	# SS-BlockedLang suprathreshold voxels	# Auditory Language Localizer suprathreshold voxels	Dice Coefficient
Left IFGorb	M(SD)= 136.2(108.42); range= 0-327	M(SD)= 130.1(112.73); range= 0-348	M(SD)= .65(0.21); range= .13-.91; *NA*=2
Left IFG	M(SD)= 199.9(134.45); range= 0-491	M(SD)= 234.9(113.30); range= 0-395	M(SD)= .72(0.19); range= .26-.91; *NA*=3
Left MFG	M(SD)= 152.3(113.16); range= 0-397	M(SD)= 95.55(85.56); range= 0-337	M(SD)= .64(0.16); range= .40-.96; *NA*=3
Left AntTemp	M(SD)= 917.5(317.11); range= 327-1304	M(SD)= 996.2(254.33); range= 362-1362	M(SD)= .82(0.11); range= .58-.96

Left PostTemp	M(SD)= 1519(505.17); range= 524-2281	M(SD)= 1431(388.66); range= 795-2384	M(SD)= .80(0.09); range= .62-.94
Left AngG	M(SD)= 81.70(92.76); range= 0-323	M(SD)= 85.85(102.84); range= 0-301	M(SD)= .39(0.24); range= 0-.80; *NA*=7

Suprathreshold voxels for each contrast, for each individual, were identified based on a threshold of z>3.09 within each language parcel.

SUPPLEMENTARY TABLE 3: SS-BlockedLang overlap in right language parcel homologues.

ROI	# SS-BlockedLang suprathreshold voxels	# Auditory Language Localizer suprathreshold voxels	Dice Coefficient
Right IFGorb	M(SD) = 74.7(90.07); range = 0-286	M(SD) = 42.1(76.86); range = 0-258	M(SD) = .51(0.27); range = 0-.90; *NA*=10
Right IFG	M(SD) = 77.6(87.50); range = 0-330	M(SD) = 93.9(119.18); range = 0-425	M(SD) = .57(0.19); range = .16-.89; *NA*= 8
Right MFG	M(SD) = 75.9(90.96); range = 0-309	M(SD) = 49.35(71.27); range = 0-198	M(SD) = .64(0.30); range = .12-.86; *NA*=11
Right AntTemp	M(SD) = 833.5(322.40); range = 236-1327	M(SD) = 848.4(272.44); range = 348-1361	M(SD) = .76(0.13); range = .36-.91
Right PostTemp	M(SD) = 855.3(489.63); range = 206-1868	M(SD) = 756.5(443.21); range = 82-1860	M(SD) = .66(0.17); range = .28-.90
Right AngG	M(SD) = 27.7(37.06); range = 0-114	M(SD) = 21.55(43.00); range = 0-136	M(SD) = .29(0.33); range = 0-.69; *NA*=16

Suprathreshold voxels for each contrast, for each individual, were identified based on a threshold of z>3.09 within each right hemisphere language parcel.

SUPPLEMENTARY TABLE 4: SS-IntDialog overlap in language parcels.

ROI	# SS-IntDialog suprathreshold voxels	# Auditory Language Localizer suprathreshold voxels	Dice Coefficient
Left IFGorb	M(SD) = 117.15 (93.68); range = 0-297	M(SD) = 130.10(112.73); range = 0-348	M(SD) = .54(0.24); range = .05-0.90; *NA*=2
Left IFG	M(SD) = 173.10(145.00); range = 35-564	M(SD) = 234.9(113.30); range = 0-395	M(SD) = 0.59(0.25); range = 0.06-0.92; *NA*=1
Left MFG	M(SD) = 174.2(95.47); range = 31-376	M(SD) = 95.55(85.56); range = 0-337	M(SD) = 0.61(0.20); range = 0.30-0.94; *NA*=3
Left AntTemp	M(SD) = 788.4(277.45); range = 387-1330	M(SD) = 996.2(254.33); range = 362-1362	M(SD) = 0.77(0.09); range = 0.57-0.92
Left PostTemp	M(SD) = 1224(390.09); range = 414-2009	M(SD) = 1431(388.66); range = 795-2384	M(SD) =0.77(0.09) ; range = 0.47-0.90
Left AngG	M(SD) = 30.15(53.72); range = 0-187	M(SD) = 85.85(102.84); range = 0-301	M(SD) = 0.29(0.28); range = 0-0.70; *NA*=11

Suprathreshold voxels for each contrast, for each individual, were identified based on a threshold of z>3.09 within each language parcel.

SUPPLEMENTARY TABLE 5: SS-IntDialog overlap in right language parcel homologues.

ROI	# SS-IntDialog suprathreshold voxels	# Auditory Language Localizer suprathreshold voxels	Dice Coefficient
Right IFGorb	M(SD) = 49.7(75.19); range = 0-231	M(SD) = 42.1(76.86); range = 0-258	M(SD) = 0.47(0.26); range = 0.14-0.72; *NA*=14
Right IFG	M(SD) = 49.60(62.50); range = 0-198	M(SD) = 93.9(119.18); range = 0-425	M(SD) = 0.33(0.14); range =0.13-0.57; *NA*=8
Right MFG	M(SD) = 48.70(69.57); range = 0-245	M(SD) = 49.35(71.27); range = 0-198	M(SD) = 0.61(0.25); range = 0.25-0.91; *NA*=11
Right AntTemp	M(SD) = 604.0(355.45); range = 66-1179	M(SD) = 848.4(272.44); range = 348-1361	M(SD) = 0.64(0.18); range = 0.18-0.90
Right PostTemp	M(SD) = 529.8(381.50); range = 11-1389	M(SD) = 756.5(443.21); range = 82-1860	M(SD) = 0.57(0.18); range = 0.24-0.85
Right AngG	M(SD) = 1.7(6.52); range = 0-29	M(SD) = 21.55(43.00); range = 0-136	*NA*=19

Suprathreshold voxels for each contrast, for each individual, were identified based on a threshold of z>3.09 within each right hemisphere language parcel.

SUPPLEMENTARY TABLE 6: SS-BlockedLang and SS-IntDialog overlap among top 100 voxels within language regions.

ROI	Dice Coefficient: SS-BlockedLang vs. Auditory Language Localizer	Dice Coefficient: SS-IntDialog vs. Auditory Language Localizer
Left IFGorb	M(SD) = .69(.16); range = .19-.91	M(SD) = .65(.14); range = .38-.87

Left IFG	M(SD) = .71(.12); range = .48-.87	M(SD) = .59(.15); range = .32-.82
Left MFG	M(SD) = .73(.16); range = .28-.94	M(SD) = .71(.13); range = .42-.93
Left AntTemp	M(SD) = .57(.13); range = .28-.77	M(SD) = .56(.16); range = .27-.87
Left PostTemp	M(SD) = .55(.17); range = .14-.81	M(SD) = .48(.18); range = .21-.76
Left AngG	M(SD) = .43(.31); range = 0-.80	M(SD) = .41(.21); range = .03-.71

Top 100 voxels were identified for Forward>Backward contrast in SS-BlockedLang and SS-IntDialog, and Intact>Degraded contrast in the Auditory Language Localizer, within each language parcel.

SUPPLEMENTARY TABLE 7: SS-IntDialog magnitude in language regions and right language parcel homologues.

ROI	Forward	Backward	Backward>Forward
Left IFGorb	M(SD) = 2.68(1.43); range = 0.54-6.14	M(SD) = 0.73(0.60); range = -0.17-2.06	Est.= -1.95 S.E.= 0.23 t-value= -8.34 p-value< .001*
Left IFG	M(SD) = 3.63(1.10); range = 1.89-5.89	M(SD) = 1.30(0.64); range = 0.39-2.38	Est.= -2.33 S.E.= 0.24 t-value= -9.71 p-value< .001*
Left MFG	M(SD) = 3.08(1.21); range = 1.35-6.52	M(SD) = 0.38(0.68); range = -0.70-1.87	Est.= -2.70 S.E.= 0.21 t-value= -12.70 p-value< .001*
Left AntTemp	M(SD) = 4.36(1.61); range = 2.49-9.00	M(SD) = 1.45(1.10); range = 0.04-4.68	Est.= -2.91 S.E.= 0.20 t-value= -14.20 p-value< .001*

Left PostTemp	M(SD) = 7.48(2.32); range = 4.65-13.84	M(SD) = 2.99(1.69); range = 0.70-7.43	Est.= -4.49 S.E.= 0.21 t-value= -21.29 p-value< .001*
Left AngG	M(SD) = 1.37(1.44); range = -1.05-4.45	M(SD) = 0.50(1.12); range = -1.21-4.10	Est.= -0.87 S.E.= 0.17 t-value= -5.07 p-value< .001*
Right IFGorb	M(SD) = 1.26(0.88); range = -0.22-2.96	M(SD) = 0.52(0.59); range = -0.55-1.58	Est.= -0.75 S.E.= 0.17 t-value= -4.30 p-value< .001*
Right IFG	M(SD) = 2.32(1.26); range = 0.11-5.03	M(SD) = 1.35(0.68); range = 0.45-3.65	Est.= -0.97 S.E.= 0.23 t-value= -4.15 p-value< .001*
Right MFG	M(SD) = 2.04(1.79); range = -0.41-6.85	M(SD) = 1.01(1.16); range = -0.49-4.76	Est.= -1.02 S.E.= 0.26 t-value= -3.98 p-value< .001*
Right AntTemp	M(SD) = 5.15(2.42); range = 0.97-12.17	M(SD) = 2.44(1.83); range = -0.01-6.65	Est.= -2.72 S.E.= 0.23 t-value= -11.87 p-value< .001*
Right PostTemp	M(SD) = 6.49(2.53); range = 2.96-11.94	M(SD) = 3.41(1.67); range = 0.52-6.99	Est.= -3.08 S.E.= 0.29 t-value= -10.57 p-value< .001*
Right AngG	M(SD) = 0.53(0.76); range = -1.10-1.71	M(SD) = 0.67(0.85); range = -1.28-2.24	Est.= 0.14 S.E.= 0.10 t-value= 1.34 p-value= 0.19

Average magnitude (betas) for SS-IntDialog conditions (Forward and Backward), extracted from ss-fROIs for language based on the auditory language localizer. Results (Est. = estimate, S.E. = standard error, t-value, and uncorrected p-value) from the model: lmer(mean_topvoxels_extracted~b_or_f +(1|participantID), REML = FALSE)
** indicates significance level p<.05, Bonferroni corrected for 6 ROIs (p<.0083)*

SUPPLEMENTARY TABLE 8: ss-fROI identification for language regions.

ROI	SS-BlockedLang	SS-IntDialog	Auditory Language Localizer
Left IFGorb	.55	.50	.50
Left IFG	.70	.55	.85
Left MFG	.60	.80	.45
Left AntTemp	1	1	1
Left PostTemp	1	1	1
Left AngG	.30	.15	.40

Proportion of participants (out of 20) who had at least 100 voxels that significantly responded to the language contrast (Forward>Backward for SS-BlockedLang and SS-IntDialog; Intact>Degraded for the auditory language localizer) at a threshold of Z>3.09.

SUPPLEMENTARY TABLE 9: SS-BlockedLang statistics in SS-BlockedLang-defined language regions.

ROI	Backward v. Forward	Dialogue v. Monologue	Interaction
Left IFGorb	Est.= -1.64 S.E.= 0.20 t-value= -8.03 p< .001 *	Est.= 0.05 S.E.= 0.20 t-value= 0.24 p= 0.813	Est.= -0.12 S.E.= 0.29 t-value= -0.43 p= 0.67
Left IFG	Est.= -1.93	Est.= 0.07	Est.= -0.04

	S.E.= 0.19 t-value= -10.29 p< .001 *	S.E.= 0.19 t-value= 0.40 p= 0.69	S.E.= 0.26 t-value= -0.16 p= 0.87
Left MFG	Est.= -1.92 S.E.= 0.21 t-value= -9.36 p< .001 *	Est.= 0.18 S.E.= 0.21 t-value= 0.86 p= 0.40	Est.= -0.11 S.E.= 0.29 t-value= -0.38 p= 0.71
Left AntTemp	Est.= -2.91 S.E.= 0.17 t-value= -17.31 p< .001 *	Est.= 0.42 S.E.= 0.17 t-value= 2.49 p= 0.02	Est.= -0.25 S.E.= 0.24 t-value= -1.07 p= 0.29
Left PostTemp	Est.= -3.86 S.E.= 0.20 t-value= -19.59 p< .001 *	Est.= 0.42 S.E.= 0.20 t-value= 2.12 p= 0.04	Est.= -0.15 S.E.= 0.28 t-value= -0.54 p= 0.59
Left AngG	Est.= -1.05 S.E.= 0.17 t-value= -6.12 p< .001 *	Est.= 0.50 S.E.= 0.17 t-value= 2.91 p= 0.005 *	Est.= -0.26 S.E.= 0.24 t-value= -1.08 p= 0.29

Differences between conditions in held-out data within each language ss-fROI defined based on SS-BlockedLang. Results (Est. = estimate, S.E. = standard error, t-value, and uncorrected p-value) from the model:
*lmer(mean_topvoxels_extracted~b_or_f*d_or_m+(1|participantID), REML = FALSE)*
** indicates significance level p<.05, Bonferroni corrected for 6 ROIs (p<.0083)*

SUPPLEMENTARY TABLE 10: SS-IntDialog statistics in SS-IntDialog-defined language regions.

ROI	Backward>Forward
Left IFGorb	Est.= -1.83; S.E.= 0.23; t-value= -7.88; **p-value <.001***
Left IFG	Est.= -2.32; S.E.= 0.23; t-value= -10.10; **p-value <.001***
Left MFG	Est.= -2.74; S.E.= 0.23; t-value= -11.75; **p-value <.001***

ROI	Backward>Forward
Left IFGorb	Est.= -1.83; S.E.= 0.23; t-value= -7.88; **p-value <.001***
Left AntTemp	Est.= -2.90; S.E.= 0.21; t-value= -13.51; **p-value <.001***
Left PostTemp	Est.= -4.18; S.E.= 0.26; t-value= -16.27; **p-value <.001***
Left AngG	Est.= -0.66; S.E.= 0.21; t-value= -3.08; **p-value= .006***

Differences between Forward and Backward conditions in held-out data within each language ss-fROI defined based on SS-IntDialog (uncorrected p-values shown). 18 participants had two runs of SS-IntDialog and were included in this analysis. Results (Est. = estimate, S.E. = standard error, t-value, and uncorrected p-value) from the model: lmer(mean_topvoxels_extracted~b_or_f +(1|participantID), REML = FALSE)
** indicates significance level p<.05, Bonferroni corrected for 6 ROIs (p<.0083)*

Additional Methods

fMRIprep specifications

**Note that the following boilerplate description is copied verbatim from the fMRIprep output, per guidelines from the developers.*

Results included in this manuscript come from preprocessing performed using fMRIPprep 1.2.6 (Esteban, Markiewicz, et al. (2018); Esteban, Blair, et al. (2018); RRID:SCR_016216), which is based on Nipype 1.1.7 (Gorgolewski et al. (2011); Gorgolewski et al. (2018); RRID:SCR_002502).

Anatomical data preprocessing

The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) using N4BiasFieldCorrection (Tustison et al. 2010, ANTs 2.2.0), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped using

antsBrainExtraction.sh (ANTs 2.2.0), using OASIS as target template. Brain surfaces were reconstructed using recon-all (FreeSurfer 6.0.1, RRID:SCR_001847, Dale, Fischl, and Sereno 1999), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle (RRID:SCR_002438, Klein et al. 2017). Spatial normalization to the ICBM 152 Nonlinear Asymmetrical template version 2009c (Fonov et al. 2009, RRID:SCR_008796) was performed through nonlinear registration with antsRegistration (ANTs 2.2.0, RRID:SCR_004757, Avants et al. 2008), using brain-extracted versions of both T1w volume and template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (FSL 5.0.9, RRID:SCR_002823, Zhang, Brady, and Smith 2001).

Functional data preprocessing

For each of the 10 BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. The BOLD reference was then co-registered to the T1w reference using bbregister(FreeSurfer) which implements boundary-based registration (Greve and Fischl 2009). Co-registration was configured with nine degrees of freedom to account for distortions remaining in the BOLD reference. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt (FSL 5.0.9, Jenkinson et al. 2002). The BOLD time-series, were resampled to surfaces on the following spaces: fsaverage5. The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying a single, composite

transform to correct for head-motion and susceptibility distortions. These resampled BOLD time-series will be referred to as preprocessed BOLD in original space, or just preprocessed BOLD. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. Automatic removal of motion artifacts using independent component analysis (ICA-AROMA, Pruim et al. 2015) was performed on the preprocessed BOLD on MNI space time-series after removal of non-steady state volumes and spatial smoothing with an isotropic, Gaussian kernel of 6mm FWHM (full-width half-maximum). Corresponding "non-aggresively" denoised runs were produced after such smoothing. Additionally, the "aggressive" noise-regressors were collected and placed in the corresponding confounds file. The BOLD time-series were resampled to MNI152NLin2009cAsym standard space, generating a preprocessed BOLD run in MNI152NLin2009cAsym space. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. Several confounding time-series were calculated based on the preprocessed BOLD: framewise displacement (FD), DVARS and three region-wise global signals. FD and DVARS are calculated for each functional run, both using their implementations in Nipype (following the definitions by Power et al. 2014). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (CompCor, Behzadi et al. 2007). Principal components are estimated after high-pass filtering the preprocessed BOLD time-series (using a discrete cosine filter with 128s cut-off) for the two CompCor variants: temporal (tCompCor) and anatomical (aCompCor). Six tCompCor components are then calculated from the top 5% variable voxels within a mask covering the subcortical regions. This subcortical mask is obtained by heavily eroding the brain mask, which ensures it does not include cortical GM regions. For aCompCor, six components are calculated within the intersection of the

aforementioned mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run (using the inverse BOLD-to-T1w transformation). The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. All resamplings can be performed with a single interpolation step by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and template spaces). Gridded (volumetric) resamplings were performed using antsApplyTransforms (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos 1964). Non-gridded (surface) resamplings were performed using mri_vol2surf(FreeSurfer).

Many internal operations of fMRIPrep use Nilearn 0.5.0 (Abraham et al. 2014, RRID:SCR_001362), mostly within the functional processing workflow. For more details of the pipeline, see the section corresponding to workflows in fMRIPrep's documentation.

References (fMRIprep)

Abraham, Alexandre, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gael Varoquaux. 2014. "Machine Learning for Neuroimaging with Scikit-Learn." Frontiers in Neuroinformatics 8. https://doi.org/10.3389/fninf.2014.00014.

Avants, B.B., C.L. Epstein, M. Grossman, and J.C. Gee. 2008. "Symmetric Diffeomorphic Image Registration with Cross-Correlation: Evaluating Automated Labeling of Elderly and Neurodegenerative Brain." Medical Image Analysis 12 (1): 26–41. https://doi.org/10.1016/j.media.2007.06.004.

Behzadi, Yashar, Khaled Restom, Joy Liau, and Thomas T. Liu. 2007. "A Component Based Noise Correction Method (CompCor) for BOLD and Perfusion Based fMRI." NeuroImage 37 (1): 90–101. https://doi.org/10.1016/j.neuroimage.2007.04.042.

Dale, Anders M., Bruce Fischl, and Martin I. Sereno. 1999. "Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction." NeuroImage 9 (2): 179–94. https://doi.org/10.1006/nimg.1998.0395.
Esteban, Oscar, Ross Blair, Christopher J. Markiewicz, Shoshana L. Berleant, Craig Moodie, Feilong Ma, Ayse Ilkay Isik, et al. 2018. "FMRIPrep." Software. Zenodo. https://doi.org/10.5281/zenodo.852659.
Esteban, Oscar, Christopher Markiewicz, Ross W Blair, Craig Moodie, Ayse Ilkay Isik, Asier Erramuzpe Aliaga, James Kent, et al. 2018. "fMRIPrep: A Robust Preprocessing Pipeline for Functional MRI." Nature Methods. https://doi.org/10.1038/s41592-018-0235-4.
Fonov, VS, AC Evans, RC McKinstry, CR Almli, and DL Collins. 2009. "Unbiased Nonlinear Average Age-Appropriate Brain Templates from Birth to Adulthood." NeuroImage, Organization for human brain mapping 2009 annual meeting, 47, Supplement 1: S102. https://doi.org/10.1016/S1053-8119(09)70884-5.
Gorgolewski, K., C. D. Burns, C. Madison, D. Clark, Y. O. Halchenko, M. L. Waskom, and S. Ghosh. 2011. "Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python." Frontiers in Neuroinformatics 5: 13. https://doi.org/10.3389/fninf.2011.00013.
Gorgolewski, Krzysztof J., Oscar Esteban, Christopher J. Markiewicz, Erik Ziegler, David Gage Ellis, Michael Philipp Notter, Dorota Jarecka, et al. 2018. "Nipype." Software. Zenodo. https://doi.org/10.5281/zenodo.596855.
Greve, Douglas N, and Bruce Fischl. 2009. "Accurate and Robust Brain Image Alignment Using Boundary-Based Registration." NeuroImage 48 (1): 63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060.
Jenkinson, Mark, Peter Bannister, Michael Brady, and Stephen Smith. 2002. "Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images." NeuroImage 17 (2): 825–41. https://doi.org/10.1006/nimg.2002.1132.
Klein, Arno, Satrajit S. Ghosh, Forrest S. Bao, Joachim Giard, Yrjö Häme, Eliezer Stavsky, Noah Lee, et al. 2017. "Mindboggling Morphometry of Human Brains." PLOS Computational Biology 13 (2): e1005350. https://doi.org/10.1371/journal.pcbi.1005350.
Lanczos, C. 1964. "Evaluation of Noisy Data." Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis 1 (1): 76–85. https://doi.org/10.1137/0701007.
Power, Jonathan D., Anish Mitra, Timothy O. Laumann, Abraham Z. Snyder, Bradley L. Schlaggar, and Steven E. Petersen. 2014. "Methods to Detect, Characterize, and Remove Motion Artifact in Resting State fMRI." NeuroImage 84 (Supplement C): 320–41. https://doi.org/10.1016/j.neuroimage.2013.08.048.

Pruim, Raimon H. R., Maarten Mennes, Daan van Rooij, Alberto Llera, Jan K. Buitelaar, and Christian F. Beckmann. 2015. "ICA-AROMA: A Robust ICA-Based Strategy for Removing Motion Artifacts from fMRI Data." NeuroImage 112 (Supplement C): 267–77. https://doi.org/10.1016/j.neuroimage.2015.02.064.
Tustison, N. J., B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee. 2010. "N4ITK: Improved N3 Bias Correction." IEEE Transactions on Medical Imaging 29 (6): 1310–20. https://doi.org/10.1109/TMI.2010.2046908.
Zhang, Y., M. Brady, and S. Smith. 2001. "Segmentation of Brain MR Images Through a Hidden Markov Random Field Model and the Expectation-Maximization Algorithm." IEEE Transactions on Medical Imaging 20 (1): 45–57. https://doi.org/10.1109/42.906424.

Task performance

For SS-BlockedLang and SS-IntDialog, we recorded button presses after each video when a still image of Elmo appeared on the screen. Accuracy on the attention checks was 99.4% overall for SS-BlockedLang and 99.2% overall for SS-IntDialog. We preregistered that we would exclude runs if the participant missed more than 50% of the attention checks; no runs were excluded based on this criterion. For the ToM localizer, we measured accuracy on the true/false questions; accuracy was 77.5% overall. We preregistered that we would exclude runs with less than 70% accuracy on the ToM localizer. However, upon examining the accuracy after data collection, we realized that this would result in the exclusion of 1 run of ToM localizer from 10 of our 20 participants. Upon inspection, many participants with lower accuracy had no responses recorded for some trials, presumably because they responded to the true/false prompt after the recording window. Because this task was being used as a localizer and we had reason to believe that participants were aiming to read the blocks and answer the questions, we decided not to exclude trials based on accuracy.

ss-fROI identification within SS-BlockedLang and SS-IntDialog tasks

We expected to be able to identify an ss-fROI in each search space using SS-IntDialog and SS-BlockedLang that we could identify using the auditory language localizer. Our ss-fROI detection method identifies the top 100 voxels that respond to a contrast within a search space, no matter what the level of response is. Thus, for the purpose of determining how well our LIT tasks compare to the auditory language localizer, we determined how many language regions had at least 100 voxels that significantly responded to the language contrast (Forward>Backward for SS-BlockedLang and SS-IntDialog; Intact>Degraded for the auditory language localizer) at a threshold of Z>3.09 (**Supplementary Table 8**). Overall, we found a similar number of people who had identifiable ss-fROIs by region for each task.

If the LIT tasks function well as language localizers, then we also expected that we would get similar results when localizing and testing held-out data using the same task, as we got from defining ss-fROIs based on the auditory language localizer. We used an in-lab script that iteratively used the z-stat image of each 3/4 (for SS-BlockedLang) or 1/2 (for SS-IntDialog) combined runs (i.e., each 'fold') to determine the top 100 voxels for a given subject, ROI, and contrast. We then used the cope image from the left-out run of a given iteration to extract the betas from these selected top voxels, then averaged these betas together per participant (**Supplementary Figures 2, 3; Supplementary Tables 9, 10**).

"And now," cried Max, "let the wild rumpus start!"

— Maurice Sendak, ***Where the Wild Things Are***

Chapter 3 : Using fMRI to study language processing in awake toddlers

**This chapter includes preliminary analyses on a small subset of toddlers who have already participated in the fMRI study. These will not be the final set of analyses submitted for publication, and thus should be interpreted accordingly.*

Abstract

Toddlers undergo immense changes in their language comprehension and production in a short period of time. However, we know quite little about the neural underpinnings of language comprehension during this important developmental period, as awake toddlers are very challenging to study using functional MRI. We developed a novel task using 20-second videos of *Sesame Street*, in which the audio stream was either played normally (Forward) or reversed by character (Backward), while the characters either spoke to the viewer (Monologue) or to each other (Dialogue). First, we confirmed that toddlers would attend to all the stimuli conditions via an online behavioral study. Next, we began scanning awake toddlers. Using the Forward>Backward contrast, we examined (1) group-level activation for the language contrast in the whole brain, (2) individual-level activation within language regions by condition, using individually-defined functional regions of interest for language iteratively defined and tested in held-out data, and (3) lateralization for language within individual participants. Preliminary results from N=6 toddlers with usable data (ages 26-36 months) suggest that we can measure language-evoked activation in canonical language regions in this age group, and that this activation may be left lateralized. Though preliminary, these results point to the possibility and promise of studying language processing in the brains of awake toddlers.

Introduction

To understand the neural basis of language development, it is critically important - but profoundly challenging - to study toddlers. Toddlerhood epitomizes language learning in action: toddlers undergo rapid advances in their language comprehension and production (Frank et al., 2021; Ganger & Brent, 2004), and they differ dramatically in their individual trajectories of language acquisition (e.g., Frank et al., 2021). This is also an important period for identifying future difficulties with language, as delays or disorders of language acquisition may be early signs of later diagnosed neurodevelopmental disorders, such as autism (Wetherby et al., 2004; Zwaigenbaum et al., 2005).

Neural measures may be particularly informative for measuring language in toddlers because explicit task performance may mis- and underestimate their language abilities (e.g., Bates, 1993; Bloom & German, 2000; Onishi & Baillargeon, 2005). However, toddlers pose a truly formidable challenge for acquiring functional MRI (fMRI) data from awake participants. Toddlers are notoriously difficult and change-averse, and fMRI scanning is demanding of participants (e.g., to pay attention while holding very still) and can be anxiety-provoking. Given these challenges, some previous studies have opted to scan toddlers during natural sleep[13]. Auditory speech elicits neural responses during sleep even in young infants (Cheour et al., 2002; Dehaene-Lambertz et al., 2002; Peña et al., 2003; Shultz et al., 2014). Prior work in sleeping toddlers found that the younger toddlers (age 2 years) had more activation in frontal, occipital, and cerebellar areas than older toddlers (age 3 years) while listening to comprehensible

[13] It should be noted that scanning sleeping toddlers is also very challenging.

speech (Redcay et al., 2008). While an important step, this work is limited by the fact that participants were merely listening to speech rather than comprehending language.

Measuring neural activity during (awake) language processing in toddlers requires experimental designs that powerfully and spontaneously engage toddlers' sustained attention, while simultaneously being highly-powered and well-controlled enough to isolate language-related evoked activity. We embarked on designing such a task by using one of the most powerful tools we could think of: Elmo.

Indeed, we designed two fMRI tasks using clips from episodes of *Sesame Street*, a television show with decades of programming explicitly designed for our target age group[14]. Particularly in pediatric neuroimaging, there has been a movement toward using naturalistic stimuli – such as commercially available movies – in order to sustain attention, decrease movement in the scanner, and evoke strong signals (Cantlon, 2020; Cantlon & Li, 2013; Kamps et al., 2022; Redcay & Moraczewski, 2020; Richardson et al., 2018; Vanderwal et al., 2019). However, a tradeoff for engagement and high data quality can be rigorous experimental control. In studies of language, for instance, a typical control condition is backward or foreign speech - which do not regularly appear in commercially available movies. To introduce this control, we created a control condition for language by reversing the audio track of characters' speech and overlaying it on the video (for more details see Chapter 2).

A second benefit of using *Sesame Street* videos for our stimuli was that we could include additional contextual information for the language excerpts. The intersection

[14] Sesame Street was also used in fMRI experiments by (Cantlon & Li, 2013; Emerson & Cantlon, 2012)

between social context and linguistic processing may be especially relevant in early childhood, as toddlers learning language rely heavily on linguistic input from their environment. This input can come from sources directed to the child (e.g., a caregiver speaking directly to the child), and also from observing other people using language to communicate with each other (e.g., observing two family members having a conversation). Social context has been shown to play a critical role in language learning (e.g., Kuhl, 2007, 2011), and social cognition also develops rapidly during early childhood (for a review, see (Soto-Icaza et al., 2015)). Thus, we included videos in which one character at a time speaks directly to the child ("monologue") and videos in which two characters engage in a conversation ("dialogue").

In conjunction with a neuroimaging study to determine whether these tasks could elicit robust and reliable language responses in adult brains (they do; see Chapter 2), we also wanted to ensure our target age group would attend to the stimuli. Thus, we conducted an online behavioral pilot study to measure toddlers' attention to different stimuli conditions. After confirming that toddlers did not preferentially attend to some conditions over others, we began scanning awake toddlers using our novel language tasks. This chapter will first detail our stimulus development procedures, then describe how we validated the stimuli in a behavioral study with toddlers, and finally conclude with preliminary results from toddlers who participated in the fMRI study thus far.

Part 1: Stimulus Development and Characterization

In designing a language localizer task for toddlers, we had to balance multiple desiderata, including: (1) engaging stimuli that would keep children in the scanner, (2)

variable but constrained content, and (3) an appropriate control condition for language.

To create an engaging task for awake 18-36-month-olds, we realized that we would need audiovisual stimuli, rather than audio-only stimuli. Movies have been shown to decrease motion and improve data quality compared to traditional resting state scans in older children (Frew et al., 2022). Drawing inspiration from other awake functional neuroimaging work in children, we decided to use naturalistic stimuli: clips from a television show or movie (Cantlon, 2020; Cantlon & Li, 2013; Kamps et al., 2022; Redcay & Moraczewski, 2020; Richardson et al., 2018; Vanderwal et al., 2019). In terms of the source material, we needed something with lots of different scenes, as we did not want to repeat stimuli and introduce familiarity effects. We also wanted to choose material that we knew would be attractive to our target age group. Thus, we decided to use video clips from *Sesame Street*: a long-running educational program designed with our target age in mind. We quickly realized that *Sesame Street* was ideal for our purposes. The scenes are filmed with live-action puppets, and are therefore somewhat constrained in their visual properties. The content of the scenes varies, but the structure of scenes and episodes is fairly consistent, meaning that it was possible to find many similar clips to reduce variability. In particular, many scenes involve either one puppet talking to the viewer (i.e., monologue) or two puppets talking to each other (i.e., dialogue). This allowed us to compare child-directed and overheard speech, a distinction which is hypothesized to impact language learning in this age group (e.g., Shneidman et al., 2013; Weisleder & Fernald, 2013).

In addition to deciding on appropriate source material, we also had to design an appropriate control condition. To control for the visual properties of the scenes, we

knew that we needed some condition that combined a language control condition with the video clips. Thus, we decided to overlay our language control-condition audio on visual scenes played normally. Because we were combining the auditory language control with the video, we were also concerned about introducing new confounds – in particular, we did not want the comprehensible language condition to differ from the control condition other than by language comprehensibility. With this in mind, we considered three choices for the control-condition audio stream that have been used in previous neuroimaging studies: foreign speech, degraded speech, and backwards speech. Foreign speech has the benefit of actually being language, and directly controls for comprehensibility specifically (e.g., Schlosser et al., 1998). We tried to find *Sesame Street* clips in other languages, but it was not possible to find enough of the exact same scenes in different languages, particularly languages that none of our participants would be familiar with. Furthermore, even when some clips were dubbed into other languages, the voice actors differed from the English versions, which could impact the interpretation of the stimuli. Another option we considered was acoustically degraded speech that controlled for lower-level auditory features (e.g., Overath et al., 2015; Scott et al., 2017; Stoppelman et al., 2013). However, in the version we tried, the acoustically degraded speech sounded much more unnatural than the normal speech overlaid on the video, and we were therefore concerned that condition effects could be due to the unnaturalness of the control condition. We therefore opted for backwards speech, specifically by reversing the audio stream of each character. This allowed us to maintain the "voice" of each character even in the control condition, and to align the backwards speech with the mouth movements of the puppets. Here again, *Sesame Street* proved to be an apt choice of source material, as the rigid mouth movements of the puppets seemed to have less of an audio-visual mismatch than we would have had with human actors. Young children are sensitive to audio-visual

mismatch in speech, and thus we did not want that to be a confound in the control condition (Gogate & Bahrick, 1998; Lewkowicz & Flom, 2014).

Having selected our stimuli, we then created two language tasks using the *Sesame Street* clips. The first – Sesame Street-Blocked Language (SS-BlockedLang) – utilizes a block design with four conditions: Forward Monologue, Forward Dialogue, Backward Monologue, and Backward Dialogue. Each condition involves clips from *Sesame Street*, with either a single character speaking to the viewer (Monologue) or two characters speaking to each other (Dialogue). The accompanying audio is either played normally (Forward), or temporally reversed by character (Backward), rendering the language incomprehensible. This approach allowed us to use rich, engaging, and social multimodal imaging, while controlling for the visual features of the stimuli. Block designs are a staple of functional neuroimaging research, as they allow for the isolation of cognitive processes using carefully designed experimental control conditions. In this case, the SS-BlockedLang task allows us to identify language processing by contrasting forward and backward speech, while controlling for the visual and contextual features of the stimuli. Forward versus backward speech has been used in many neuroimaging studies to isolate language comprehension (e.g., Bedny et al., 2011; Dehaene-Lambertz et al., 2002; Moore-Parks et al., 2010; Redcay et al., 2008).

A concern, however, is that block designs in neuroimaging research rely on the *only* differences in the neural signal between blocks being attributable to the experimenter-controlled differences between conditions. Unintentional features of the stimuli that differ between conditions, or unaccounted for differences in individual responses to the conditions (such as attention and interest, see Chapter 4), can impact the interpretation of the results. In this case, we were particularly concerned that if toddlers did not

attend to the backwards condition as much as the forwards condition, they may move more in the scanner during backwards blocks. Thus, we also designed a second task (Sesame Street-Interleaved Dialogue, or SS-IntDialog) which incorporated full scenes (1-3 minutes each) in which two characters engaged in a dialogue. We reversed the audio for one of the two characters for the duration of the scene, such that forward and backward speech alternated for each clip. In this way, the length of the backwards segments would be shorter than the 20-second blocks in the SS-BlockedLang task, and potentially less likely to be impacted by attention differences. We created stimuli for both tasks and tested them behaviorally in toddlers, and using fMRI in adults (Chapter 2).

Development

SS-BlockedLang: We first selected candidate *Sesame Street* clips for our stimuli set. Monologue clips were trimmed to 10 seconds and contained only one puppet that spoke in English directly to the viewer for the full length of the clip. Clips were not cut off in the middle of a word or clause, and generally did not include: distracting noises (unless they can be trimmed out of the clip), discussions of negative emotions (e.g., feeling sad or angry), singing, excessive laughter, solely counting, or certain characters that used particularly atypical language (e.g., Cookie Monster). Dialogue clips were trimmed to 20 seconds and followed the same criteria, except they had to contain exactly two characters that spoke to each other. Once candidate clips were selected, Backward versions of each clip were created by reversing the speech stream, by character, for each clip. The audio was normalized for each clip. Each candidate clip was then evaluated by a second experimenter, who checked that these criteria were met and additionally excluded clips in which the Backward speech did not align well

with the puppet's mouth movements. Final clips were selected, keeping in mind balancing characters. Two 10-second monologue clips were combined to form each 20-second monologue block. Stimuli were split into two sets, with Forward and Backward versions of each clip in different sets (such that an individual would never see two versions of the same clip).

SS-IntDialog: The criteria for stimulus selection were the same as for SS-BlockedLang, but in this case, we selected for entire scenes rather than short segments. For each scene, we created two versions: one in which character A's audio was played forward and character B's audio was reversed, and one in which character A's audio was reversed and character A's audio was played forward. Other sounds in the clips were played forward. Stimuli were split into two sets, so that participants only saw a given scene once. For the behavioral pilot only, we also created 90-second versions of the clips so that length did not differ between conditions, and we additionally included all-forward clips and all-backwards clips.

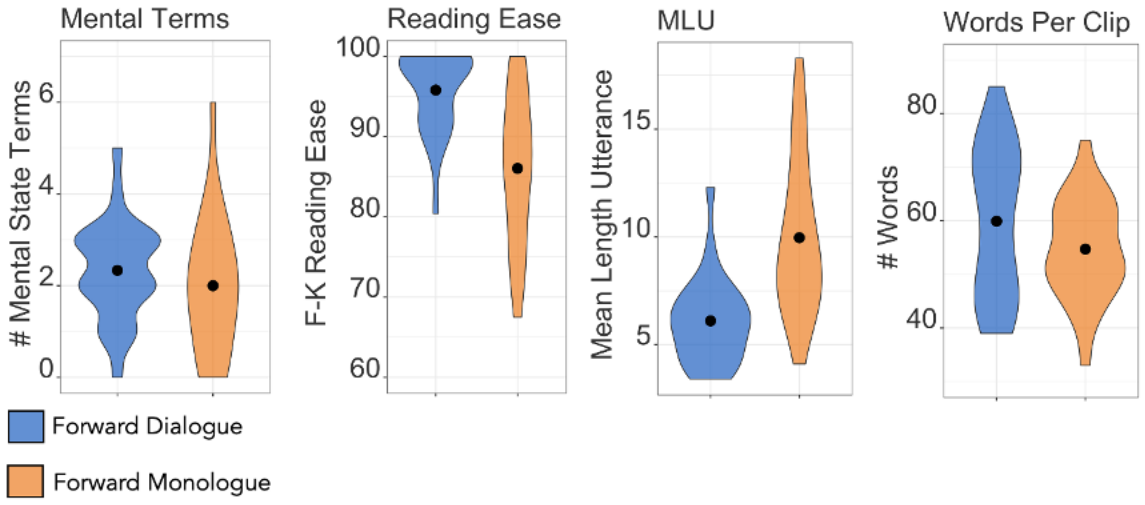
Characterization

To characterize the final stimulus set, we transcribed each video and analyzed key linguistic features[15]. We focused on stimuli features in SS-BlockedLang, as we wanted to determine what linguistic differences were present between the monologue and dialogue blocks. Mental state terms were identified by an expert in this domain as terms that expressed mental states and emotions; we counted the total number of mental state terms per block. Reading ease was calculated using the Flesch reading

[15] Transcriptions and key features can be found on OSF: https://osf.io/whsb7/

ease formula, which takes into account words per sentence and syllables per word, resulting in a value between 0 to 100 with 100 being the easiest to read (Flesch, 1948). Mean length of utterance (MLU) was calculated as the average number of words per sentence for each block, and Words Per Clip was the total number of words per block.

Figure 3.1: SS-BlockedLang stimuli characterization.



Average number of mental terms, Flesch reading ease score, mean length of utterance, and total number of words per clip for the dialogue (blue) and monologue (orange) stimuli in the SS-BlockedLang task.

Using two sample t-tests, there was no difference between dialogue and monologue for mental state terms (t=.89, p=.38) or total number of words per block (t=1.5, p=.14). Flesch reading ease score was higher for dialogue than monologue (t=4.48, p<.001), and monologues had more words per sentence (t=-4.56, p<.001). While these results are matched by some features (e.g., mental state terms, total words), the differences in complexity could plausibly contribute to condition differences between monologue and dialogue. This was not the case in adults (see Chapter 2), but could still plausibly impact language processing in young children for whom more complex language

might require additional resources to process. If complexity differences did drive condition differences in toddlers' language-evoked activation, we might imagine higher responses to monologue than dialogue in language regions.

Summary

We created two novel tasks using *Sesame Street* video clips. First, in both tasks, we were able to vary whether the language was comprehensible or not, allowing us to isolate and evoke a 'language response.' Second, in our SS-BlockedLang task, we were also able to vary whether the language occurred in dialogue (i.e., two characters talking to each other) or monologue (i.e., one character at a time talking to the viewer). While there are some differences in the linguistic features between the monologue and dialogue - in particular, the monologue clips use slightly more complex language - the conditions are fairly well-matched. The next step was to determine whether toddlers would watch them.

Part 2: Behavioral Pilot

The purpose of the online pilot study was to measure toddlers' attention to the *Sesame Street* stimuli in order to inform the fMRI study design. Specifically, we wanted to ensure toddlers equally attended to videos of forward and backward speech, as a difference in attention could otherwise impact our ability to compare conditions in the neuroimaging portion of the study. We created two experimental designs that we tested behaviorally in toddlers. The first was a block design (SS-BlockedLang), which included 20-second segments of forward and backward speech. While this design would maximize power to detect a difference in brain activity, a concern was that toddlers would not tolerate 20-second blocks of backwards speech. Thus, the second

design (SS-IntDialog) alternated between forward and backward speech in shorter intervals: with speaker changes in a natural dialogue.

Methods

Participants. 51 children (ages 14-34 months, mean(SD)=24(4.5) months) participated in the behavioral pilot. 36 participants were tested using SS-BlockedLang, and 15 participants were tested using SS-IntDialog. Families were recruited to our general participant database through a variety of advertising approaches, such as Facebook ads. Families provided basic demographic information about their child and contact information. We reached out to eligible participants via email and invited them to sign up for a time slot to participate in the study via an online scheduler. We recruited participants between the ages of 12-40 months who were born at >37 weeks gestation. A given child only participated in one version of the experiment (either SS-BlockedLang or SS-IntDialog). Parents were sent a copy of the consent form prior to the study, and consent was obtained verbally before the experiment began with the experimenter. This study was approved by the MIT Committee on the Use of Humans as Experimental Subjects.

SS-BlockedLang Experiment: Our first-choice experimental design to use with toddlers was a 2x2 block design varying (1) comprehensibility of speech, and (2) whether the speech was directed to the viewer or was a conversation between two characters. To manipulate speech comprehensibility, we varied whether the auditory stream for each speaker was played forward or backward. To manipulate the type of speech, we presented 20-second clips of dialogue (two characters, simultaneously present, alternately speaking) or two sequential 10-second clips of monologue (one character,

present alone, followed by a second character). Thus, we have four conditions: forward monologue, backward monologue, forward dialogue, and backward dialogue. Forward and backward versions of each clip were counterbalanced between participants (randomly assigned Set A or Set B). Trials were presented in a balanced, pseudorandom order. Order of trial conditions was also counterbalanced between participants, such that each version began with a different condition. There were 4 orders, and two sets of videos, leading to a total of 8 different experiment versions. There were a total of 12 trials per condition (48 total trials). Children continued the experiment until they watched all 48 trials or decided to stop (approximately 17 minutes total).

SS-IntDialog Experiment: Our second-choice experimental design for the toddler fMRI study was interleaved forward and backward speech. Participants watched 90-second clips of dialogue (two characters, simultaneously present, alternately speaking) in three conditions: forward (audio streams for both characters were played forward), backward (audio streams for both characters were played backward), and alternating (audio stream for one character was played forward and the other was played backward). Note that we cut the videos to include the first 90s of the scene so that the total duration would not confound the results (though in the fMRI SS-IntDialog task, the entire scenes were played). The alternating versions of the clips were counterbalanced between participants (e.g., one child saw Elmo forward and Abby backward for a clip where they blow bubbles; another child saw Elmo backward and Abby forward for that same clip). Trials were presented in a balanced, pseudorandom order. Order of trial conditions was also counterbalanced between participants, such that each version began with a different condition. There were 6 orders, and two sets of videos, leading to a total of 12 different experiment versions. There were 2 trials per condition (6 total

trials). Children continued the experiment until they watched all 6 trials or decided to stop (approximately 10 minutes total).

Data Collection: Data collection was conducted over Zoom with a live experimenter. The session was recorded and stored securely for offline coding. For SS-BlockedLang, an unlisted YouTube playlist (whichever one the participant was randomly assigned) was shared with the caregiver, who was instructed to pull up the playlist on their device and share their screen and audio. When this was unsuccessful (e.g., sometimes sharing the screen made the participant's own video disappear), the experimenter pulled up the videos and shared their screen and audio. For SS-IntDialog, the experimenter shared video and audio for the stimuli videos (we found that this worked better when transitioning to using SS-IntDialog due to challenges uploading videos to YouTube). The caregiver was instructed not to direct the child's attention back to the screen during the videos, and the caregiver and child were told to let the experimenter know if the child was "all done." The experimenter also sometimes ended the experiment if the child seemed upset or was no longer paying any attention to the experiment. The child was instructed to look at the screen at the beginning of the experiment (e.g., "Look, CHILD! Here we go!"). An attention-getter appeared on the screen in the four corners to help experimenters determine where the edges of the child's screen were for coding of looking behavior. An attention getter appeared after every trial in the center of the screen. Participants were compensated with a $5 Amazon gift card after the Zoom session. They were also sent a unique survey link to fill out the MacArthur-Bates Communicative Development Inventory (available at webcdi.stanford.edu; (Fenson et al., 2006)). We used the Words and Sentences form, which is designed and normed for 16–30-month-olds. Participants were compensated an additional $5 once

they completed the survey. Surveys had to be completed within two weeks of study participation.

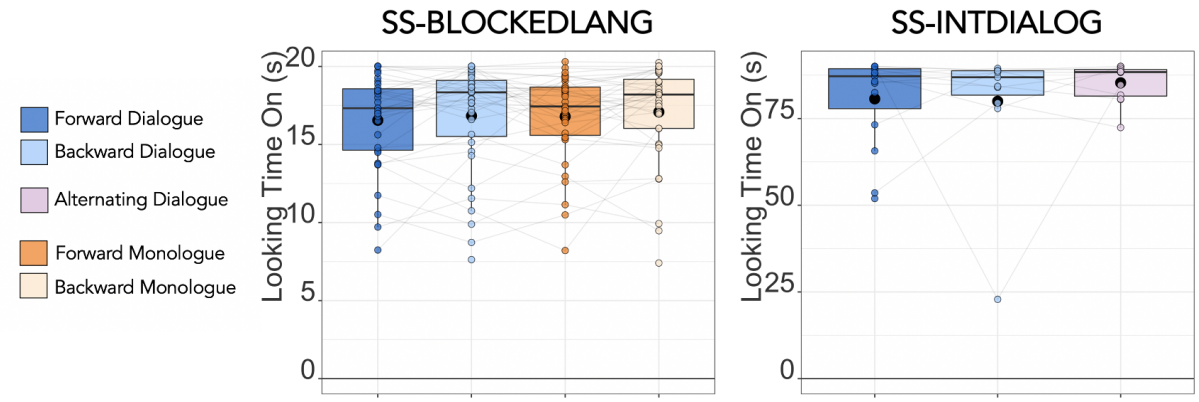
Data Processing and Analysis: Looking time was coded offline using DataVyu (Datavyu Team, 2014). A member of the research team first identified the timestamp for the start of each trial based on the end of the attention getter. Looks were coded as "on" or "off" depending on whether the child was looking at the screen. Sometimes, it was impossible to tell if the child was looking on or off the screen due to their face being out of the camera view; these trials were marked as invalid. For these exploratory analyses, we only included the valid trials in which we could determine whether the child was looking at the screen or not. In SS-BlockedLang, the maximum looking time per trial was 20s, and in SS-IntDialog, the maximum looking time per trial was 90s. For each child, we calculated the average looking time for all the trials they watched. We then conducted one-way repeated measures ANOVAs to test for condition differences.

Results

We found no differences in looking time between conditions for either version of the task in this exploratory pilot sample. For SS-BlockedLang, participants had 3-48 valid trials (mean(SD)=29.58(13.24) trials) out of a possible 48 trials. Note that we did not expect children to watch all 17 minutes of our stimuli, and we stopped presenting the videos when children fussed out. A one-way repeated measures ANOVA was conducted to examine the effect of condition on average looking time for all valid trials per condition per participant; looking time did not differ between conditions (Figure 3.2; $F(3, 102)=2.109$, $p=.10$). Furthermore, for the trials children watched, average looking time per condition per child was close to the 20s trial duration (range=7.4-

20.3s, mean(SD)=16.8(3.1) s). For SS-IntDialog, participants had 1-6 valid trials (mean(SD)=4.87(1.77) trials) out of a possible 6 trials. Looking time again did not differ between conditions (F(2, 21)=.63, p=.54). Even though these trials were much longer than the SS-BlockedLang trials, average looking time per condition per child was close to the 90s trial duration (range=22.9-90 s, mean(SD)=81.9(13.5) s).

Figure 3.2: No differences in looking time between conditions.



There were no differences in looking time for valid trials between conditions in either the 20-second blocks for forward and backward monologue and dialogue (left) or the 90-second dialogue clips with forward, backward, or alternating speech (right).

Summary

This online pilot study suggested that toddlers in our intended age range would (at least sometimes) watch the videos we created. Critically, there was no effect of condition on looking time in either experiment. We had been concerned that children may prefer the forward to the backward videos, which could create a confound in an fMRI experiment. However, since this was not the case, we proceeded with the stimuli for the fMRI study.

Part 3: fMRI Study

The goal of this study is to collect functional data from awake toddlers, in order to measure the brain's response during language comprehension. Data collection is ongoing. Preliminary results from the current sample are provided below.

Methods

Participants: We recruited and tested toddlers between the ages of 18-36 months. 6 toddlers (ages 26-36 months, mean(SD)= 31.7(3.5) months, 4 female) provided usable functional MRI data and are included in these preliminary analyses. 3 toddlers with functional data were excluded for not having at least one usable run, and an additional 19 toddlers participated in the study but did not complete any functional scans (4 of these children began some anatomical scanning). It is worth noting that most attempted scans did not result in data collection: of 44 scan sessions across 28 children, 18 sessions involved some data collection, and 10 sessions involved functional data collection.

Experimental Protocol and Pediatric Neuroimaging Considerations: The protocol was approved by the MIT Committee on the Use of Humans as Experimental Subjects. Informed consent was provided by parents of participants. Participants were compensated at a rate of $100/session, and $5/survey for filling out the MB-CDI (Fenson et al., 2006).

We employed numerous strategies to acquire fMRI data from toddlers, many inspired by prior research (Raschle et al., 2012; Richardson et al., 2018; Thieba et al., 2018). In

addition to the stimuli-specific considerations, we also developed various techniques to try with individual children.

Framing: We introduced the MRI scanner to children as a "rocket ship," and framed the visit as a "trip to space." This framing has been used in a number of pediatric neuroimaging settings, as it provides a fun, plausible explanation for getting into a narrow tube (the "rocket ship"), wearing a head coil ("space helmet"), and hearing background sounds during image acquisition ("noises while blasting through space"). We decorated the scanner with space stickers and used bedsheets with rocket ships.

Pre-Visit Preparations: Parents were sent detailed instructions before the visit to help them know what to expect. At least one week prior to the visit, parents were asked to start preparing their child by: (1) having their child watch a video we created of another child participating in the study[16], (2) reading a book to their child about Elmo going to space in a rocket ship, that outlined the steps they would follow during the visit[17], (3) playing scanner noises to acclimate their child to the sounds, and (4) practicing listening to favorite songs through in-ear headphones. We told parents that they could bring along (metal-free) favorite stuffed animals and pajamas to help their child feel as comfortable as possible. Parents were also asked to send along favorite videos and songs so that we could play them while getting children set up in the scanner.

---

[16] https://youtu.be/P2__yGWnp7s

[17] https://www.google.com/url?q=https://drive.google.com/file/d/1ck77qhOjt2S6VKKnBLF7vAax3QTpe-iB/view?usp%3Dsharing&sa=D&source=editors&ust=1679146685883657&usg=AOvVaw0O86zq1HheWTwQqfWV7sd5

Visit Protocol: Parents and children met the lead experimenter and the secondary experimenter in the playroom, which contained a mock scanner setup. While the lead experimenter acquired informed consent from the parent and answered questions, the secondary experimenter played with the child to establish familiarity and rapport. This was important, as the secondary experimenter would be in the scanner room with the child during data acquisition. Next, as much time as needed was taken to get the child comfortable and ready to go into the "real rocket ship." This could include: reading the "Elmo Goes to Space" book, picking out stuffed animals to come with them into the scanner, practicing putting their stuffed animals in the mock scanner, climbing into the mock scanner themselves, eating a snack, diaper changes, measuring height and weight, taking off shoes, and checking for metal using a handheld metal detector ("magic wand"). Parents changed into metal-free scrubs if they wished to accompany their child in the scanner room, and a second metal detector check for parents and children was completed before entering the scanner control room. To encourage children to enter the scanner room, we let them pick out a space sticker to stick on the "rocket ship." Then, the goal was to put in the in-ear headphones for hearing protection, secure the headphones in place with an EarBandIt headband on top, get the child to lay down, place the mirror on top, place the top of the headcoil (we snuck this under the mirror so that the child was already watching their videos), then move the child to isocenter in the scanner. These steps were incredibly difficult for this population. Our strategies included: playing favorite songs in the headphones[18] ("listen, what song do you hear?"), playing favorite videos on the projected screen at the back of the scanner ("if you lay down and look through this mirror, you can see garbage trucks!"), modeling with stuffed animals ("what does Elmo need next?"),

[18] Most popular was "Baby Shark" - thankfully (?) there is a 1-hour version on YouTube.

modeling with humans ("it's mom's turn!"), and exploring the environment ("let's press the button to make the bed go up and down"). Oftentimes children would get partway through and fuss out, but we always tried to end on a "success" (e.g., putting in headphones, laying down). In these cases, we always invited parents to come back to try again, and discussed additional strategies to help prepare their child in the interim. If the child did make it into the scanner, as much data were collected as tolerated by the child. The lead experimenter ran the scanner from the control room, using hand signals to communicate with the secondary experimenter, who was in the scanner room monitoring the child and ensuring the hearing protection remained in place. Parents were either in the scanner room or in the control room. Any time a functional task was not being run (e.g., during transitions and anatomical scans), we played child-friendly videos or videos recommended by the parents. At the end of the scan, children selected a toy to take home and were given a t-shirt.

Post-Visit: Parents completed the online version of the MB-CDI (Fenson et al., 2006) within two weeks after the visit.

Data Acquisition: Data were acquired from a 3-Tesla Siemens Magnetom Prisma scanner located at the Athinoula A. Martinos Imaging Center at MIT using a 32-channel head coil. T1-weighted structural images were acquired in 176 interleaved sagittal slices with 1.0mm isotropic voxels (MPRAGE; TA=5:53; TR=2530.0ms; FOV=256mm; GRAPPA parallel imaging, acceleration factor of 2). Functional data were acquired with a gradient-echo EPI sequence sensitive to Blood Oxygenation Level Dependent (BOLD) contrast in 3mm isotropic voxels in 46 interleaved near-axial slices covering the whole brain (EPI factor=70; TR=2s; TE=30.0ms; flip angle=90 degrees; FOV=210mm). fMRI tasks were run from a MacBook Pro laptop and projected onto a 16"x12" screen.

Participants viewed the stimuli through a mirror attached to the head coil. Tasks were run through PsychoPy software (Peirce, 2007).

fMRI Tasks:

(1) SS-BlockedLang: The priority during the scan was to collect functional data from the SS-BlockedLang task, for as many runs as possible. There were a few modifications relative to the adult version of the task. First, toddlers were not given an in-scanner button box, so there was no attention check or response period. Second, rather than using a black screen with a fixation cross during baseline, we used clips from the Inscapes video, which was explicitly designed to improve compliance of young children during neuroimaging (Vanderwal et al., 2015). These clips involved professionally-produced abstract visuals along with an instrumental soundtrack, and were designed as an alternative to resting state scans using a static fixation cross. Third, the longest fixation blocks were 10 seconds, rather than the 22 second blocks we used in the adult study. Finally, we added a clip from Inscapes into the task at the beginning, prior to the task trigger, so that toddlers would not have to wait for the scanner for something to be playing on the screen. Participants completed up to 4 runs, each approximately 5 minutes long. Each run contained unique clips, and participants never saw a Forward and Backward version of the same clip. Each run contained 3 sets of 4 blocks, one of each condition (total of 12 blocks), with 10-second rest blocks after each set of 4 blocks. Forward and Backward versions of each clip were counterbalanced between participants (randomly assigned Set A or Set B).

(2) SS-IntDialog: If toddlers were still content after finishing all of the available SS-BlockedLang runs, we attempted to run the SS-IntDialog task. The same modifications described above were made to the toddler versions of the task relative to the adult

version. Participants could complete up to 2 runs, each approximately 8 minutes long. Each run contained unique clips, and participants never saw two versions of the same clip. Each run contained 3 dialogue scenes, with 10-second rest blocks between each. Versions of each clip were counterbalanced between participants (randomly assigned Set A or Set B).

fMRI Preprocessing and Statistical Modeling: FMRI data were first preprocessed using fMRIPrep 22.0.2 (Esteban et al., 2019), which is based on Nipype 1.8.5 (Gorgolewski et al., 2011). See Supplementary Materials for full preprocessing pipeline details. We used a lab-specific script that uses Nipype to combine tools from several different software packages for first-level modeling. Each event regressor was defined as a boxcar convolved with a standard double-gamma HRF, and a high-pass filter (1/210 Hz) was applied to both the data and the model. Artifact detection was performed using Nipype's RapidART toolbox (an implementation of SPM's ART toolbox). Individual TRs were marked as outliers if (1) there was more than 1 unit of frame displacement, or (2) the average signal intensity of that volume was more than 3 standard deviations away from the mean average signal intensity. We included one regressor per outlier volume. In addition, we included a summary movement regressor (framewise displacement) and 6 anatomical CompCor regressors to control for the average signal in white matter and CSF. We applied a 6mm smoothing kernel to preprocessed BOLD images. The first-level model was run using FSL's GLM in MNI space. Subject level modeling was performed with in-lab scripts using Nipype. Specifically, FSL's fixed effects flow was used to combine runs at the level of individual participants. A subject level model was created for each set of usable runs per contrast for each task (up to 4 runs for SS-BlockedLang). Runs with more than 33% of timepoints marked as outliers, or runs without data from all four conditions, were excluded from analysis. Output average

magnitudes in each voxel in the second level model were then passed to the group level model. Group modeling used in-lab scripts that implemented FSL's RANDOMISE to perform a nonparametric one-sample t-test of the con values against 0 (5000 permutations, MNI space, threshold alpha = .05), accounting for familywise error. For preliminary analyses, we show the uncorrected group random effects analyses at a lenient threshold, since there is no significant activation at the corrected threshold.

Whole Brain Analysis: For preliminary group whole brain results, we used a threshold of t>1.5, uncorrected, given that only 6 subjects were included. For language comprehension, we used the [Forward Dialogue + Forward Monologue] > [Backward Dialogue + Backward Monologue] contrast.

Functional Region of Interest Analysis: 4 participants had more than 1 usable run of the SS-BlockedLang task and were included in fROI analyses. We iteratively defined subject-specific functional regions of interest (ss-fROIs) for language as the top 100 voxels activated in an individual, within each of five predefined language search spaces, for the Forward>Backward contrast in n-1 runs, where n is the total usable runs (Fedorenko et al., 2010). The five language search spaces in the left hemisphere included: Left IFGorb, Left IFG, Left MFG, Left AntTemp, and Left PostTemp (similar to Fedorenko et al, 2010; in this case, 5 parcels in left hemisphere were created based on a group-level probabilistic activation overlap map for a sentences>nonwords contrast in 220 adult participants[19]; we excluded the Left AngG parcel based on the adult data, Chapter 2). We also looked within the mirror of these search spaces in the right hemisphere (i.e., right hemisphere language homologues). We used an in-lab script

[19] Parcels downloaded from https://evlab.mit.edu/funcloc/

that iteratively used the z-stat image of each n-1/n combined runs (i.e., each 'fold') to determine the top 100 voxels for a given subject and ROI. We then used the cope image from the left-out run of a given iteration to extract the betas from these selected top voxels, then averaged these betas together per participant.

Lateralization: Another common measure of language in the brain is the laterality index; that is, how many more voxels are activated for language on the left compared to the right. We calculated the laterality index using the formula, $LI=(V_{left}-V_{right}+1)/(V_{left}+V_{right}+2)$, where V refers to the number of suprathreshold voxels (Berl et al., 2014; Desmond et al., 1995). For preliminary analyses, we calculated LI using a threshold of z>1.68 and a cluster threshold of k=10 voxels, constrained to the left language search spaces and their mirrored right hemisphere homologue search spaces. This is a more lenient threshold than we preregistered, given the lack of suprathreshold voxels at the preregistered level of p<.01.

Preregistration: Hypotheses and analysis plan were preregistered on OSF[20], including a plan to look at preliminary data for this thesis. Given a smaller sample of usable data than we anticipated, and smaller effects than we anticipated in this initial sample, we are using more lenient threshold than the ones we preregistered.

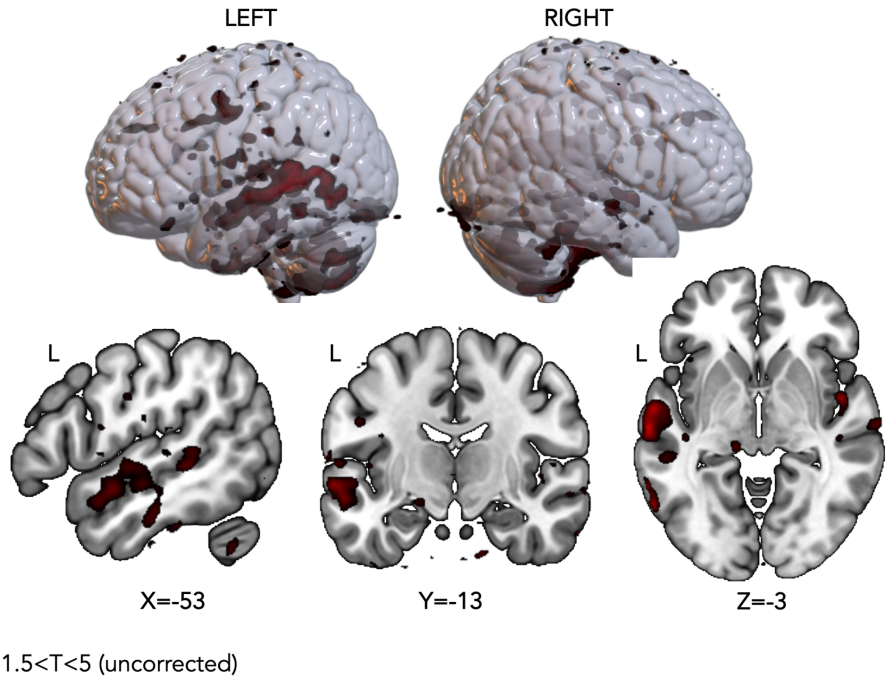
[20] https://osf.io/wvqht

Preliminary Results

Whole-brain group effects

As a first step to determine whether there were hints of language-evoked activation in toddlers using this task, we conducted a group whole-brain analysis including the 6 participants with at least one usable run of the functional task. At an uncorrected, low threshold, we see some left temporal and possible left frontal activation for the language contrast (Forward>Backward; **Figure 3.3**).

Figure 3.3: Group whole-brain random effects analysis for language.

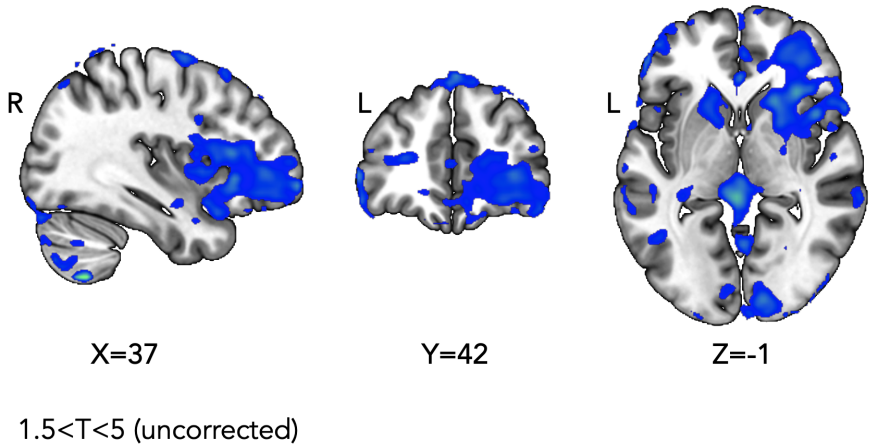


Uncorrected t-map for the language contrast (Forward>Backward), including data from 6 participants. Visualized at an uncorrected threshold t-value=1.5.

We also visualized the Backward Dialogue>Backward Monologue contrast at the group level, again at an uncorrected, low threshold, as a measure of which regions may

respond more to two characters interacting compared to one character on the screen (**Figure 3.4**). As a sanity check, we see that different regions are evoked for this contrast than the language contrast, such as visual cortex and right prefrontal cortex.

Figure 3.4: Group whole-brain random effects analysis for two interacting characters.

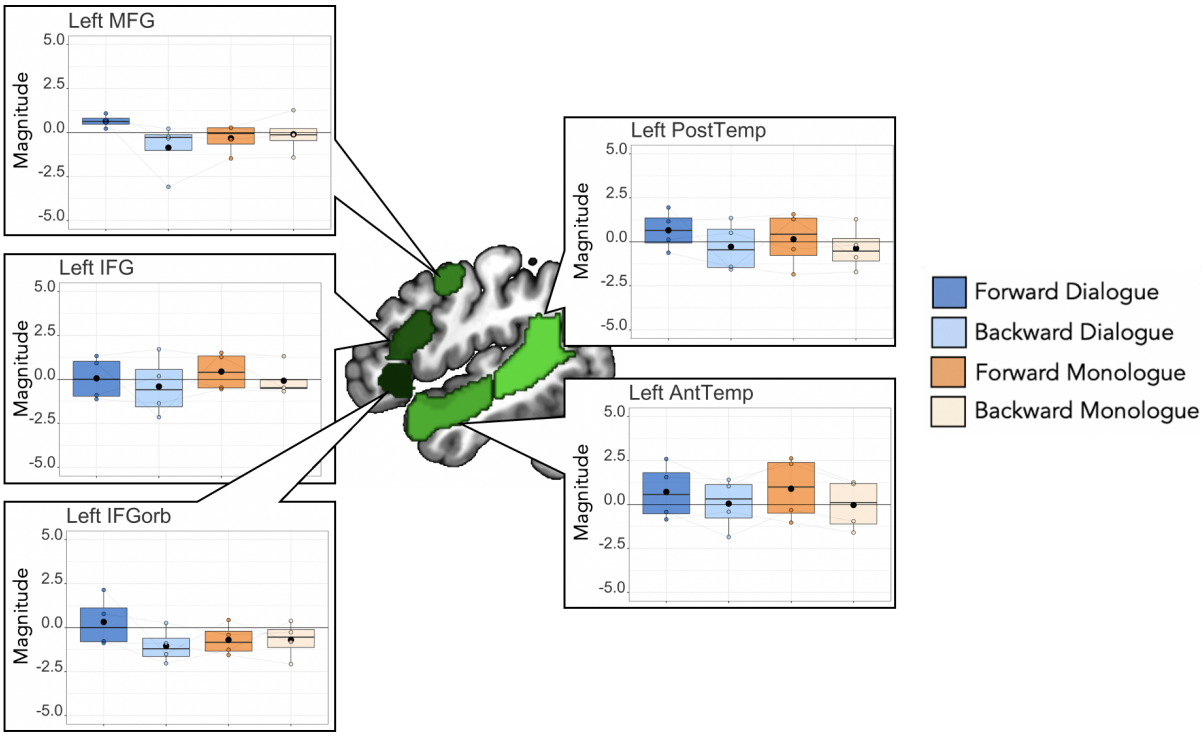


Uncorrected t-map for the Backward Dialogue>Backward Monologue contrast, including data from 6 participants. Visualized at an uncorrected threshold t-value=1.5.

Univariate responses to task conditions within language regions

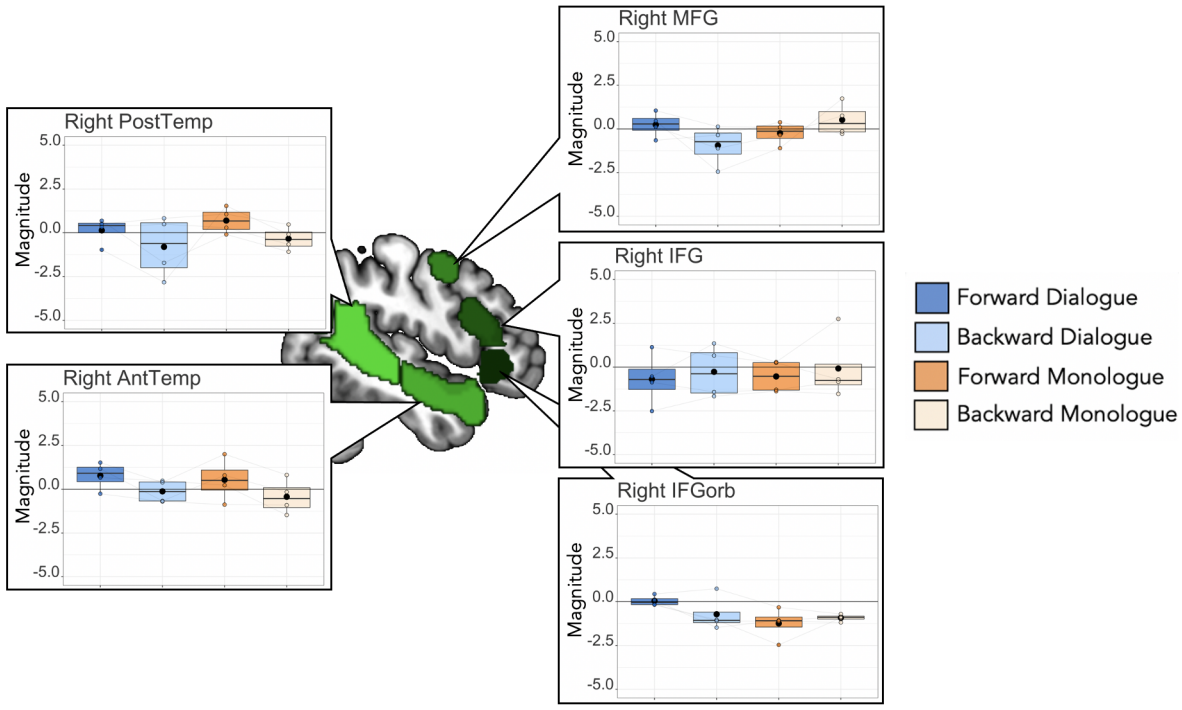
Next, we examined the response to each condition within language fROIs. Average response magnitude for the top-100 "language" voxels per region (defined by the Forward>Backward contrast) were iteratively extracted from held-out runs (**Figure 3.5**). Although the sample size is much too small to make any inferences at this point, there are promising hints of higher responses to forward than backward speech in language regions. We additionally extracted responses from right hemisphere homologues of language regions (**Figure 3.6**). There are again preliminary hints of higher responses to forward than backward speech in right temporal regions.

Figure 3.5: Univariate responses per condition in language fROIs.



***Center**: Left hemisphere language parcels overlaid on template brain (green; parcels include left IFGorb, IFG, MFG, AntTemp, and PostTemp from https://evlab.mit.edu/funcloc/). **Panels**: Average response magnitude (betas) per individual for each condition in the SS-BlockedLang task was extracted from subject-specific functional regions of interest for language, defined by the Forward>Backward contrast in independent data (blue: Forward Dialogue; light blue: Backward Dialogue; orange: Forward Monologue; light orange: Backward Monologue). Boxplot with mean in black circle; colored circles show individual participants with light gray lines connecting single participants.*

Figure 3.6: Univariate responses per condition in right hemisphere homologues of language regions.



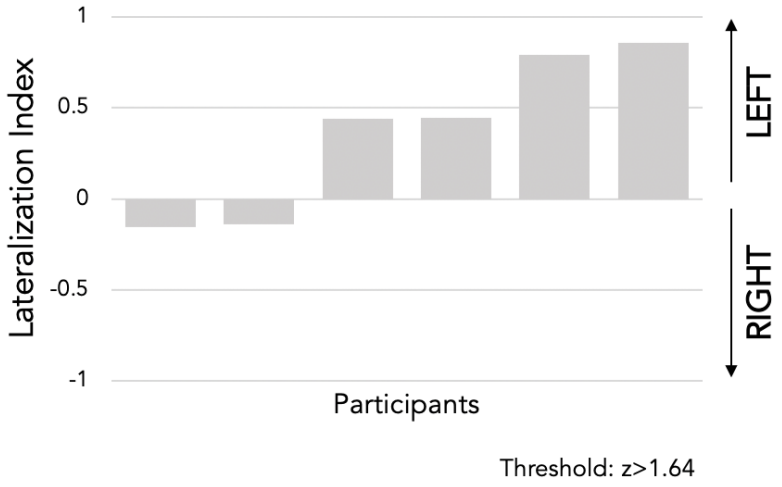
***Center**: Right hemisphere homologues of language parcels overlaid on template brain (green; parcels include right IFGorb, IFG, MFG, AntTemp, and PostTemp from https://evlab.mit.edu/funcloc/). **Panels**: Average response magnitude (betas) per individual for each condition in the SS-BlockedLang task was extracted from subject-specific functional regions of interest for language, defined by the Forward>Backward contrast in independent data (blue: Forward Dialogue; light blue: Backward Dialogue; orange: Forward Monologue; light orange: Backward Monologue). Boxplot with mean in black circle; colored circles show individual participants with light gray lines connecting single participants.*

Lateralization for language response

Finally, to determine whether language activation in toddlers is lateralized to the left hemisphere, we calculated lateralization index within left hemisphere language search spaces and mirror search spaces in the right hemisphere. We used a more lenient threshold than planned (z>1.64, cluster threshold k=10) because multiple participants

had few to no suprathreshold voxels at the more stringent planned threshold. While preliminary, there was some evidence of left lateralization for language (**Figure 3.7**).

Figure 3.7: Lateralization for language within language search spaces.



Lateralization index for 6 participants with usable functional data. Lateralization index was calculated using the formula LI=(V_{left}-V_{right}+1)/(V_{left}+V_{right}+2), where V is the number of suprathreshold voxels (here, z>1.64, cluster correction k=10) within left hemisphere language search spaces (V_{left}: sum of suprathreshold voxels in left IFGorb, IFG, MFG, AntTemp, and PostTemp) and right hemisphere language search spaces (V_{right}: sum of suprathreshold voxels in right IFGorb, IFG, MFG, AntTemp, and PostTemp).

Summary

Preliminary results from 6 toddlers with usable functional data suggest that it is possible to measure language-evoked activation using the SS-BlockedLang task. Across all preliminary measures, there were hints of a response to Forward>Backward speech in the expected left-lateralized canonical language regions.

Discussion

This chapter in some ways functions as a proof-of-concept for scanning awake toddlers on a language task. The first hurdle was creating the task itself: figuring out how to create an fMRI task that would be engaging for very young children, but also afforded us some degree of experimental control in order to extract a language response. The second check was to make sure toddlers would watch the stimuli, and in particular, to make sure that there were no attentional differences between conditions that might impede our ability to measure reliable responses with fMRI. Finally, the biggest challenge was to figure out how to get toddlers in the scanner, and to determine whether we could measure any language-evoked activation in their brains. On all fronts, this chapter suggests that we are on the right track.

The use of commercially-produced content designed for children was inspired by multiple neuroimaging studies in children (Cantlon, 2020; Cantlon & Li, 2013; Kamps et al., 2022; Redcay & Moraczewski, 2020; Richardson et al., 2018; Vanderwal et al., 2019). Children move less in the scanner when they are watching movies (Frew et al., 2022), and multiple studies have shown that it is possible to extract robust and reliable responses for cognitive processes like face processing and theory of mind (Kamps et al., 2022; Richardson et al., 2018). A particular challenge we encountered was that a good control for language does not occur naturally – thus, we decided to reverse the speech stream by character. Because even young children are sensitive to the correspondence between auditory and visual cues in speech (Gogate & Bahrick, 1998; Lewkowicz & Flom, 2014), we decided to use clips of puppets with rigid mouth movements. From our own perspectives, the edited clips 'looked like' the characters were speaking. Given that children did not show condition preferences in their looking

time behavior, and that adult responses recapitulated the expected language-evoked activation patterns in the brain (Chapter 2), we feel fairly confident in our stimuli design.

When it came to scanning toddlers, our protocol was largely inspired by previous fMRI research with infants (Kosakowski et al., 2022) and children (Richardson et al., 2018). Preparation was crucial with this age group – the more time spent preparing for the visit (e.g., by watching our video, or practicing with headphones), the better things went, generally. Flexibility and willingness to adapt to the needs of the child were also critical. Inspired by one of my other studies (Chapter 4), we asked caregivers to send their child's favorite videos and songs, so that we could play them as we got the child in the scanner. This served as an incentive to lay down and look at the screen, and also provided distraction and comfort as the child was moved into the scanner. Particularly for pediatric populations, tailoring the experience to the individual may not only be beneficial, but also quite important (see Chapter 4).

Though the fMRI data are very preliminary, the hints so far suggest that these data may eventually speak to important debates in the literature regarding the development of language-selective regions in cortex. For example, while language processing seems to be left lateralized even shortly after birth (Peña et al., 2003), some evidence suggests that language processing becomes more left-lateralized with time (Berl et al., 2014; Holland et al., 2007), potentially through decreased activation in the right hemisphere (Martin et al., 2022; Olulade et al., 2020). This debate is missing a critical period of time, however – is language network differently lateralized in toddlers whose language skills are rapidly increasing? Another question is whether temporal and frontal regions show different trajectories of involvement in language processing during development.

For instance, there is some evidence in children that temporal regions become strongly and stably left-lateralized before frontal regions (Berl et al., 2014). Similarly, in sleeping toddlers, prior work suggested that speech processing in 2-year-olds may recruit additional regions outside the canonical language network, that are not as engaged in 3-year-olds (Redcay et al., 2008). Is this the case for language comprehension in awake toddlers? Finally, in our own adult data, we found that language regions did not differentiate between dialogue and monologue speech (Chapter 2). However, we know that young children may be especially sensitive to the social context of language exposure (e.g., Hoff, 2006; Kuhl, 2007). Thus, do language regions in toddlers differentiate between child-directed versus overheard language? While our preliminary data cannot yet speak to these important questions, we are optimistic that a complete dataset might be able to do so.

Limitations

The results from the fMRI study thus far are very preliminary: with only 6 toddlers, we can only cautiously peek at the data. In particular, the results visualized here use more lenient thresholds than we preregistered, as the more stringent thresholds resulted in few to no suprathreshold voxels. This may be an indication that we need more data, or it may be an indication that the effects we are seeing are not robust. Furthermore, we opted to stick as close as possible to the same fMRI processing pipeline that we used in the adult study (Chapter 2). Toddler data is much sparser and messier than adult data, however, and we may find in further analyses that some analytical choices impact the results. To account for this, we have preregistered a number of robustness checks that we will perform and report with the full dataset. Finally, it is important to note that most children we tried to test did not even get in the scanner. The participants who are

included here are, for the most part, at the older end of the age range we attempted scanning and were fairly verbal, which may limit the generalizability of future results.

Conclusions

These preliminary results suggest that it is possible to measure brain activation during language comprehension in awake toddlers using fMRI. Intriguingly, there are hints that this activation may be left lateralized and present in canonical language regions.

Acknowledgments

This research was supported by the Simons Foundation Autism Research Initiative via the Simons Center for the Social Brain at MIT and the NSF Graduate Research Fellowship Program (#1745302 to HO). Thank you to Somaia Saba, Hana Ro, and Michelle Hung for their assistance with stimuli creation, to Somaia Saba and Hana Ro for collecting online data with toddlers, and to Somaia Saba, Rebecca O'Connor, and Sofia Riskin for helping with toddler scanning. Thank you to Emily Chen for help with running the study and data analysis, and to Kirsten Lydic for help with fMRI data analysis. Thank you to Steve Shannon and Atsushi Takahashi from the Athinoula A. Martinos Imaging Center at MIT for supporting toddler scanning. Finally, thank you to the participating families for making this research possible.

References

Bates, E. (1993). Comprehension and Production in Early Language Development. *Monographs of the Society for Research in Child Development*, *58*(3–4), 222–242. https://doi.org/10.1111/j.1540-5834.1993.tb00403.x

Bedny, M., Pascual-Leone, A., Dodell-Feder, D., Fedorenko, E., & Saxe, R. (2011). Language processing in the occipital cortex of congenitally blind adults. *Proceedings of the National Academy of Sciences*, *108*(11), 4429–4434. https://doi.org/10.1073/pnas.1014818108

Berl, M. M., Mayo, J., Parks, E. N., Rosenberger, L. R., VanMeter, J., Ratner, N. B., Vaidya, C. J., & Gaillard, W. D. (2014). Regional differences in the

developmental trajectory of lateralization of the language network. *Human Brain Mapping*, *35*(1), 270–284. https://doi.org/10.1002/hbm.22179

Bloom, P., & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, *77*(1), B25–B31. https://doi.org/10.1016/S0010-0277(00)00096-2

Cantlon, J. F. (2020). The balance of rigor and reality in developmental neuroscience. *NeuroImage*, *216*, 116464. https://doi.org/10.1016/j.neuroimage.2019.116464

Cantlon, J. F., & Li, R. (2013). Neural Activity during Natural Viewing of Sesame Street Statistically Predicts Test Scores in Early Childhood. *PLOS Biology*, *11*(1), e1001462. https://doi.org/10.1371/journal.pbio.1001462

Cheour, M., Martynova, O., Näätänen, R., Erkkola, R., Sillanpää, M., Kero, P., Raz, A., Kaipio, M.-L., Hiltunen, J., Aaltonen, O., Savela, J., & Hämäläinen, H. (2002). Speech sounds learned by sleeping newborns. *Nature*, *415*(6872), Article 6872. https://doi.org/10.1038/415599b

Datavyu Team. (2014). *Datavyu: A Video Coding Tool*. Databrary Project, New York University. http://datavyu.org

Dehaene-Lambertz, G., Dehaene, S., & Hertz-Pannier, L. (2002). Functional Neuroimaging of Speech Perception in Infants. *Science*, *298*(5600), 2013–2015. https://doi.org/10.1126/science.1077066

Desmond, J. E., Sum, J. M., Wagner, A. D., Demb, J. B., Shear, P. K., Glover, G. H., Gabrieli, J. D. E., & Morrell, M. J. (1995). Functional MRI measurement of language Lateralization in Wada-tested patients. *Brain*, *118*(6), 1411–1419. https://doi.org/10.1093/brain/118.6.1411

Emerson, R. W., & Cantlon, J. F. (2012). Early math achievement and functional connectivity in the fronto-parietal network. *Developmental Cognitive Neuroscience*, *2*, S139–S151. https://doi.org/10.1016/j.dcn.2011.11.003

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), Article 1. https://doi.org/10.1038/s41592-018-0235-4

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2006). *MacArthur-Bates Communicative Development Inventories, Second Edition*.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233. https://doi.org/10.1037/h0057532

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and Consistency in Early Language Learning: The Wordbank Project*. MIT Press.

Frew, S., Samara, A., Shearer, H., Eilbott, J., & Vanderwal, T. (2022). Getting the nod: Pediatric head motion in a transdiagnostic sample during movie- and resting-state fMRI. *PLOS ONE*, *17*(4), e0265112. https://doi.org/10.1371/journal.pone.0265112

Ganger, J., & Brent, M. R. (2004). Reexamining the Vocabulary Spurt. *Developmental Psychology*, *40*, 621–632. https://doi.org/10.1037/0012-1649.40.4.621

Gogate, L. J., & Bahrick, L. E. (1998). Intersensory Redundancy Facilitates Learning of Arbitrary Relations between Vowel Sounds and Objects in Seven-Month-Old Infants. *Journal of Experimental Child Psychology*, *69*(2), 133–149. https://doi.org/10.1006/jecp.1998.2438

Gorgolewski, K., Burns, C., Madison, C., Clark, D., Halchenko, Y., Waskom, M., & Ghosh, S. (2011). Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. *Frontiers in Neuroinformatics*, *5*. https://www.frontiersin.org/articles/10.3389/fninf.2011.00013

Hoff, E. (2006). How social contexts support and shape language development. *Developmental Review*, *26*(1), 55–88. https://doi.org/10.1016/j.dr.2005.11.002

Holland, S. K., Vannest, J., Mecoli, M., Jacola, L. M., Tillema, J.-M., Karunanayaka, P. R., Schmithorst, V. J., Yuan, W., Plante, E., & Byars, A. W. (2007). Functional MRI of language lateralization during development in children. *International Journal of Audiology*, *46*(9), 533–551. https://doi.org/10.1080/14992020701448994

Kamps, F. S., Richardson, H., Murty, N. A. R., Kanwisher, N., & Saxe, R. (2022). Using child-friendly movie stimuli to study the development of face, place, and object regions from age 3 to 12 years. *Human Brain Mapping*, *43*(9), 2782–2800. https://doi.org/10.1002/hbm.25815

Kosakowski, H. L., Cohen, M. A., Takahashi, A., Keil, B., Kanwisher, N., & Saxe, R. (2022). Selective responses to faces, scenes, and bodies in the ventral visual pathway of infants. *Current Biology*, *32*(2), 265-274.e5. https://doi.org/10.1016/j.cub.2021.10.064

Kuhl, P. K. (2007). Is speech learning 'gated' by the social brain? *Developmental Science*, *10*(1), 110–120. https://doi.org/10.1111/j.1467-7687.2007.00572.x

Kuhl, P. K. (2011). *Social Mechanisms in Early Language Acquisition: Understanding Integrated Brain Systems Supporting Language*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195342161.013.0043

Lewkowicz, D. J., & Flom, R. (2014). The Audiovisual Temporal Binding Window Narrows in Early Childhood. *Child Development*, *85*(2), 685–694. https://doi.org/10.1111/cdev.12142

Martin, K. C., Seydell-Greenwald, A., Berl, M. M., Gaillard, W. D., Turkeltaub, P. E., & Newport, E. L. (2022). A Weak Shadow of Early Life Language Processing

Persists in the Right Hemisphere of the Mature Brain. *Neurobiology of Language*, *3*(3), 364–385. https://doi.org/10.1162/nol_a_00069

Moore-Parks, E. N., Burns, E. L., Bazzill, R., Levy, S., Posada, V., & Müller, R.-A. (2010). An fMRI study of sentence-embedded lexical-semantic decision in children and adults. *Brain and Language*, *114*(2), 90–100. https://doi.org/10.1016/j.bandl.2010.03.009

Olulade, O. A., Seydell-Greenwald, A., Chambers, C. E., Turkeltaub, P. E., Dromerick, A. W., Berl, M. M., Gaillard, W. D., & Newport, E. L. (2020). The neural basis of language development: Changes in lateralization over age. *Proceedings of the National Academy of Sciences*, *117*(38), 23477–23483. https://doi.org/10.1073/pnas.1905590117

Onishi, K. H., & Baillargeon, R. (2005). Do 15-Month-Old Infants Understand False Beliefs? *Science*, *308*(5719), 255–258. https://doi.org/10.1126/science.1107621

Overath, T., McDermott, J. H., Zarate, J. M., & Poeppel, D. (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nature Neuroscience*, *18*(6), Article 6. https://doi.org/10.1038/nn.4021

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Peña, M., Maki, A., Kovac□ić, D., Dehaene-Lambertz, G., Koizumi, H., Bouquet, F., & Mehler, J. (2003). Sounds and silence: An optical topography study of language recognition at birth. *Proceedings of the National Academy of Sciences*, *100*(20), 11702–11705. https://doi.org/10.1073/pnas.1934290100

Raschle, N., Zuk, J., Ortiz-Mantilla, S., Sliva, D. D., Franceschi, A., Grant, P. E., Benasich, A. A., & Gaab, N. (2012). Pediatric neuroimaging in early childhood and infancy: Challenges and practical guidelines. *Annals of the New York Academy of Sciences*, *1252*(1), 43–50. https://doi.org/10.1111/j.1749-6632.2012.06457.x

Redcay, E., Haist, F., & Courchesne, E. (2008). Functional neuroimaging of speech perception during a pivotal period in language acquisition. *Developmental Science*, *11*(2), 237–252. https://doi.org/10.1111/j.1467-7687.2008.00674.x

Redcay, E., & Moraczewski, D. (2020). Social cognition in context: A naturalistic imaging approach. *NeuroImage*, *216*, 116392. https://doi.org/10.1016/j.neuroimage.2019.116392

Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., & Saxe, R. (2018). Development of the social brain from age three to twelve years. *Nature Communications*, *9*(1), Article 1. https://doi.org/10.1038/s41467-018-03399-2

Schlosser, M. J., Aoyagi, N., Fulbright, R. K., Gore, J. C., & McCarthy, G. (1998). Functional MRI studies of auditory comprehension. *Human Brain Mapping*, *6*(1), 1–13. https://doi.org/10.1002/(SICI)1097-0193(1998)6:1<1::AID-HBM1>3.0.CO;2-7

Scott, T. L., Gallée, J., & Fedorenko, E. (2017). A new fun and robust version of an fMRI localizer for the frontotemporal language system. *Cognitive Neuroscience*, *8*(3), 167–176. https://doi.org/10.1080/17588928.2016.1201466

Shneidman, L. A., Arroyo, M. E., Levine, S. C., & Goldin-Meadow, S. (2013). What counts as effective input for word learning?*. *Journal of Child Language*, *40*(3), 672–686. https://doi.org/10.1017/S0305000912000141

Shultz, S., Vouloumanos, A., Bennett, R. H., & Pelphrey, K. (2014). Neural specialization for speech in the first months of life. *Developmental Science*, *17*(5), 766–774. https://doi.org/10.1111/desc.12151

Soto-Icaza, P., Aboitiz, F., & Billeke, P. (2015). Development of social skills in children: Neural and behavioral evidence for the elaboration of cognitive models. *Frontiers in Neuroscience*, *9*. https://www.frontiersin.org/articles/10.3389/fnins.2015.00333

Stoppelman, N., Harpaz, T., & Ben-Shachar, M. (2013). Do not throw out the baby with the bath water: Choosing an effective baseline for a functional localizer of speech processing. *Brain and Behavior*, *3*(3), 211–222. https://doi.org/10.1002/brb3.129

Thieba, C., Frayne, A., Walton, M., Mah, A., Benischek, A., Dewey, D., & Lebel, C. (2018). Factors Associated With Successful MRI Scanning in Unsedated Young Children. *Frontiers in Pediatrics*, *6*. https://www.frontiersin.org/articles/10.3389/fped.2018.00146

Vanderwal, T., Eilbott, J., & Castellanos, F. X. (2019). Movies in the magnet: Naturalistic paradigms in developmental functional neuroimaging. *Developmental Cognitive Neuroscience*, *36*, 100600. https://doi.org/10.1016/j.dcn.2018.10.004

Vanderwal, T., Kelly, C., Eilbott, J., Mayes, L. C., & Castellanos, F. X. (2015). Inscapes: A movie paradigm to improve compliance in functional magnetic resonance imaging. *NeuroImage*, *122*, 222–232. https://doi.org/10.1016/j.neuroimage.2015.07.069

Weisleder, A., & Fernald, A. (2013). Talking to Children Matters: Early Language Experience Strengthens Processing and Builds Vocabulary. *Psychological Science*, *24*(11), 2143–2152. https://doi.org/10.1177/0956797613488145

Wetherby, A. M., Woods, J., Allen, L., Cleary, J., Dickinson, H., & Lord, C. (2004). *Early Indicators of Autism Spectrum Disorders in the Second Year of Life*. https://doi.org/10.1007/s10803-004-2544-y

Zwaigenbaum, L., Bryson, S., Rogers, T., Roberts, W., Brian, J., & Szatmari, P. (2005). Behavioral manifestations of autism in the first year of life. *International Journal of Developmental Neuroscience*, *23*(2), 143–152. https://doi.org/10.1016/j.ijdevneu.2004.05.001

Supplementary

Preprocessing Pipeline Details

Text below is automatically generated by fMRIprep.

Results included in this manuscript come from preprocessing performed using fMRIPrep 22.0.2 (Esteban, Markiewicz, et al. (2018); Esteban, Blair, et al. (2018); RRID:SCR_016216), which is based on Nipype 1.8.5 (K. Gorgolewski et al. (2011); K. J. Gorgolewski et al. (2018); RRID:SCR_002502).

Anatomical data preprocessing

The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) with N4BiasFieldCorrection (Tustison et al. 2010), distributed with ANTs 2.3.3 (Avants et al. 2008, RRID:SCR_004757), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped with a Nipype implementation of the antsBrainExtraction.sh workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (FSL 6.0.5.1:57b01774, RRID:SCR_002823, Zhang, Brady, and Smith 2001). Brain surfaces were reconstructed using recon-all (FreeSurfer 7.2.0, RRID:SCR_001847, Dale, Fischl, and Sereno 1999), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle (RRID:SCR_002438, Klein et al. 2017). Volume-based spatial normalization to two standard spaces (MNI152NLin6Asym, MNI152NLin2009cAsym) was performed through nonlinear registration with antsRegistration (ANTs 2.3.3), using brain-extracted versions of both T1w reference and the T1w template. The following templates were selected for spatial

normalization: FSL's MNI ICBM 152 non-linear 6th Generation Asymmetric Average Brain Stereotaxic Registration Model [Evans et al. (2012), RRID:SCR_002823; TemplateFlow ID: MNI152NLin6Asym], ICBM 152 Nonlinear Asymmetrical template version 2009c [Fonov et al. (2009), RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym].

Functional data preprocessing

For each of the BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt (FSL 6.0.5.1:57b01774, Jenkinson et al. 2002). BOLD runs were slice-time corrected to 0.946s (0.5 of slice acquisition range 0s-1.89s) using 3dTshift from AFNI (Cox and Hyde 1997, RRID:SCR_005927). The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying the transforms to correct for head-motion. These resampled BOLD time-series will be referred to as preprocessed BOLD in original space, or just preprocessed BOLD. The BOLD reference was then co-registered to the T1w reference using bbregister (FreeSurfer) which implements boundary-based registration (Greve and Fischl 2009). Co-registration was configured with six degrees of freedom. Several confounding time-series were calculated based on the preprocessed BOLD: framewise displacement (FD), DVARS and three region-wise global signals. FD was computed using two formulations following Power (absolute sum of relative motions, Power et al. (2014)) and Jenkinson (relative root mean square displacement between affines, Jenkinson et al. (2002)). FD and DVARS are calculated for each

functional run, both using their implementations in Nipype (following the definitions by Power et al. 2014). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (CompCor, Behzadi et al. 2007). Principal components are estimated after high-pass filtering the preprocessed BOLD time-series (using a discrete cosine filter with 128s cut-off) for the two CompCor variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components are then calculated from the top 2% variable voxels within the brain mask. For aCompCor, three probabilistic masks (CSF, WM and combined CSF+WM) are generated in anatomical space. The implementation differs from that of Behzadi et al. in that instead of eroding the masks by 2 pixels on BOLD space, a mask of pixels that likely contain a volume fraction of GM is subtracted from the aCompCor masks. This mask is obtained by dilating a GM mask extracted from the FreeSurfer's aseg segmentation, and it ensures components are not extracted from voxels containing a minimal fraction of GM. Finally, these masks are resampled into BOLD space and binarized by thresholding at 0.99 (as in the original implementation). Components are also calculated separately within the WM and CSF masks. For each CompCor decomposition, the k components with the largest singular values are retained, such that the retained components' time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each (Satterthwaite et al. 2013). Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardized DVARS were annotated as motion outliers. Additional nuisance timeseries are calculated by means of principal components

analysis of the signal found within a thin band (crown) of voxels around the edge of the brain, as proposed by (Patriat, Reynolds, and Birn 2017). The BOLD time-series were resampled into standard space, generating a preprocessed BOLD run in MNI152NLin6Asym space. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. The BOLD time-series were resampled onto the following surfaces (FreeSurfer reconstruction nomenclature): fsaverage. Automatic removal of motion artifacts using independent component analysis (ICA-AROMA, Pruim et al. 2015) was performed on the preprocessed BOLD on MNI space time-series after removal of non-steady state volumes and spatial smoothing with an isotropic, Gaussian kernel of 6mm FWHM (full-width half-maximum). Corresponding "non-aggresively" denoised runs were produced after such smoothing. Additionally, the "aggressive" noise-regressors were collected and placed in the corresponding confounds file. Grayordinates files (Glasser et al. 2013) containing 91k samples were also generated using the highest-resolution fsaverage as intermediate standardized surface space. All resamplings can be performed with a single interpolation step by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using antsApplyTransforms (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos 1964). Non-gridded (surface) resamplings were performed using mri_vol2surf (FreeSurfer).

Many internal operations of fMRIPrep use Nilearn 0.9.1 (Abraham et al. 2014, RRID:SCR_001362), mostly within the functional processing workflow. For more details of the pipeline, see the section corresponding to workflows in fMRIPrep's documentation.

References (fMRIprep)

Abraham, Alexandre, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gael Varoquaux. 2014. "Machine Learning for Neuroimaging with Scikit-Learn." Frontiers in Neuroinformatics 8. https://doi.org/10.3389/fninf.2014.00014.

Avants, B. B., C. L. Epstein, M. Grossman, and J. C. Gee. 2008. "Symmetric Diffeomorphic Image Registration with Cross-Correlation: Evaluating Automated Labeling of Elderly and Neurodegenerative Brain." Medical Image Analysis 12 (1): 26–41. https://doi.org/10.1016/j.media.2007.06.004.

Behzadi, Yashar, Khaled Restom, Joy Liau, and Thomas T. Liu. 2007. "A Component Based Noise Correction Method (CompCor) for BOLD and Perfusion Based fMRI." NeuroImage 37 (1): 90–101. https://doi.org/10.1016/j.neuroimage.2007.04.042.

Cox, Robert W., and James S. Hyde. 1997. "Software Tools for Analysis and Visualization of fMRI Data." NMR in Biomedicine 10 (4-5): 171–78. https://doi.org/10.1002/(SICI)1099-1492(199706/08)10:4/5<171::AID-NBM453>3.0.CO;2-L.

Dale, Anders M., Bruce Fischl, and Martin I. Sereno. 1999. "Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction." NeuroImage 9 (2): 179–94. https://doi.org/10.1006/nimg.1998.0395.

Esteban, Oscar, Ross Blair, Christopher J. Markiewicz, Shoshana L. Berleant, Craig Moodie, Feilong Ma, Ayse Ilkay Isik, et al. 2018. "fMRIPrep 22.0.2." Software. https://doi.org/10.5281/zenodo.852659.

Esteban, Oscar, Christopher Markiewicz, Ross W Blair, Craig Moodie, Ayse Ilkay Isik, Asier Erramuzpe Aliaga, James Kent, et al. 2018. "fMRIPrep: A Robust Preprocessing Pipeline for Functional MRI." Nature Methods. https://doi.org/10.1038/s41592-018-0235-4.

Evans, AC, AL Janke, DL Collins, and S Baillet. 2012. "Brain Templates and Atlases." NeuroImage 62 (2): 911–22. https://doi.org/10.1016/j.neuroimage.2012.01.024.

Fonov, VS, AC Evans, RC McKinstry, CR Almli, and DL Collins. 2009. "Unbiased Nonlinear Average Age-Appropriate Brain Templates from Birth to Adulthood." NeuroImage 47, Supplement 1: S102. https://doi.org/10.1016/S1053-8119(09)70884-5.

Glasser, Matthew F., Stamatios N. Sotiropoulos, J. Anthony Wilson, Timothy S. Coalson, Bruce Fischl, Jesper L. Andersson, Junqian Xu, et al. 2013. "The Minimal Preprocessing Pipelines for the Human Connectome Project." NeuroImage, Mapping the connectome, 80: 105–24. https://doi.org/10.1016/j.neuroimage.2013.04.127.

Gorgolewski, K., C. D. Burns, C. Madison, D. Clark, Y. O. Halchenko, M. L. Waskom, and S. Ghosh. 2011. "Nipype: A Flexible, Lightweight and Extensible Neuroimaging

Data Processing Framework in Python." Frontiers in Neuroinformatics 5: 13. https://doi.org/10.3389/fninf.2011.00013.
Gorgolewski, Krzysztof J., Oscar Esteban, Christopher J. Markiewicz, Erik Ziegler, David Gage Ellis, Michael Philipp Notter, Dorota Jarecka, et al. 2018. "Nipype." Software. https://doi.org/10.5281/zenodo.596855.
Greve, Douglas N, and Bruce Fischl. 2009. "Accurate and Robust Brain Image Alignment Using Boundary-Based Registration." NeuroImage 48 (1): 63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060.
Jenkinson, Mark, Peter Bannister, Michael Brady, and Stephen Smith. 2002. "Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images." NeuroImage 17 (2): 825–41. https://doi.org/10.1006/nimg.2002.1132.
Klein, Arno, Satrajit S. Ghosh, Forrest S. Bao, Joachim Giard, Yrjö Häme, Eliezer Stavsky, Noah Lee, et al. 2017. "Mindboggling Morphometry of Human Brains." PLOS Computational Biology 13 (2): e1005350. https://doi.org/10.1371/journal.pcbi.1005350.
Lanczos, C. 1964. "Evaluation of Noisy Data." Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis 1 (1): 76–85. https://doi.org/10.1137/0701007.
Patriat, Rémi, Richard C. Reynolds, and Rasmus M. Birn. 2017. "An Improved Model of Motion-Related Signal Changes in fMRI." NeuroImage 144, Part A (January): 74–82. https://doi.org/10.1016/j.neuroimage.2016.08.051.
Power, Jonathan D., Anish Mitra, Timothy O. Laumann, Abraham Z. Snyder, Bradley L. Schlaggar, and Steven E. Petersen. 2014. "Methods to Detect, Characterize, and Remove Motion Artifact in Resting State fMRI." NeuroImage 84 (Supplement C): 320–41. https://doi.org/10.1016/j.neuroimage.2013.08.048.
Pruim, Raimon H. R., Maarten Mennes, Daan van Rooij, Alberto Llera, Jan K. Buitelaar, and Christian F. Beckmann. 2015. "ICA-AROMA: A Robust ICA-Based Strategy for Removing Motion Artifacts from fMRI Data." NeuroImage 112 (Supplement C): 267–77. https://doi.org/10.1016/j.neuroimage.2015.02.064.
Satterthwaite, Theodore D., Mark A. Elliott, Raphael T. Gerraty, Kosha Ruparel, James Loughead, Monica E. Calkins, Simon B. Eickhoff, et al. 2013. "An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data." NeuroImage 64 (1): 240–56. https://doi.org/10.1016/j.neuroimage.2012.08.052.
Tustison, N. J., B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee. 2010. "N4itk: Improved N3 Bias Correction." IEEE Transactions on Medical Imaging 29 (6): 1310–20. https://doi.org/10.1109/TMI.2010.2046908.

Zhang, Y., M. Brady, and S. Smith. 2001. “Segmentation of Brain MR Images Through a Hidden Markov Random Field Model and the Expectation-Maximization Algorithm.” IEEE Transactions on Medical Imaging 20 (1): 45–57. https://doi.org/10.1109/42.906424.

"Of course they needed to care. It was the meaning of everything."

— Lois Lowry, ***The Giver***

Chapter 4 : Personal interests amplify engagement of language regions in the brains of children with and without autism

**A version of this chapter has been submitted for publication as:*

†D'Mello, A. M., **†Olson, H. A.**, †Johnson, K. T., Nishith, S., Frosch, I. R., & Gabrieli, J. D. E. Personal interests amplify engagement of language regions in the brains of children with and without autism.
†*Authors share joint first authorship.*
Preprint: https://www.biorxiv.org/content/10.1101/2023.03.21.533695v1

Abstract

Behavioral investigations have found that personal interests can profoundly influence language-relevant behaviors; however, the influence of personal interest on language processing in the brain is unknown. We measured brain activation via functional magnetic resonance imaging (fMRI) in 20 children while they listened to personalized narratives written about their specific interests, as well as to non-personalized narratives about a neutral topic. Multiple cortical language regions, as well as select cortical and subcortical regions associated with reward and salience, exhibited greater activation for personally-interesting than neutral narratives. There was also more overlap in activation patterns across individuals for their personally-interesting narratives than neutral narratives, despite the personalized narratives being unique to each individual. These results replicated in a group of 15 children with autism, a condition characterized by both specific interests and difficulties with communication, suggesting that personally-interesting narratives may impact neural language processing even amidst challenges with language and social communication. These findings reveal that engagement with topics that are personally interesting can significantly affect activation in the neocortical and subcortical regions that subserve language, reward, and salience in the brains of children.

Introduction

Human language is informed by our personal experiences, backgrounds, intrinsic motivations, and interests. However, when studying language processing in the laboratory, researchers typically use impersonal and generic stimuli with the assumption that idiosyncrasies and personal relevance merely introduce noise (Van Lancker, 1991). Crucially, failing to consider the effects of individual differences in interest may affect brain activation in unknown ways and potentially obscure some of the functionality of the language network.

Personal interests can be powerful motivators of language comprehension and communication (Krapp, Hidi, & Renninger, 1992). Interesting materials increase reading comprehension performance, allowing children to better comprehend materials beyond their established reading level (e.g., Shnayer, 1968), and children are more likely to play with and be generous towards individuals who share similar interests (e.g., Sparks, Schinkel, & Moore, 2017). Perhaps most strikingly, interest can improve performance in populations that typically struggle with language. For example, case studies of children with autism, a condition characterized by communication difficulties as well as a high prevalence of specific interests (Klin, Danovitch, Merz, & Volkmar, 2007), find positive impacts of scaffolding sociolinguistic interactions around topics of personal interest (e.g., Harrop, Amsbary, Towner-Wright, Reichow, & Boyd, 2019). Notably, few studies have extended personal interests into the brain. This may be in part due to a reluctance to use idiosyncratic stimuli and thereby give up experimental control. Prior studies have, however, personalized stimuli to the individual when studying certain phenomena – such as food cravings or memories – based on the

intuition that personalization (e.g., a favorite food, or a video of a specific memory) might be the most effective and ecological way to elicit neural responses (e.g., Bainbridge & Baker, 2022; Tomova et al., 2020).

Despite the effects of interest on linguistically-relevant behavior, and the intuitive use of personalization to study brain activation in other domains, no studies have examined how topics of personal interest modulate language activation in the human brain. We recruited children (n=20, 6.98-12.01 years) with highly specific interests for an individually-tailored functional magnetic resonance imaging (fMRI) experiment in which they listened to personalized narratives written about their interests. We compared brain responses to these narratives with responses to non-personalized, control narratives about nature that were the same across all children. We hypothesized that personally-interesting narratives would elicit higher activation than neutral narratives in language regions. We also explored whether personal interests would affect language network function in a group of children with autism (n=15, 8.18-13.27 years).

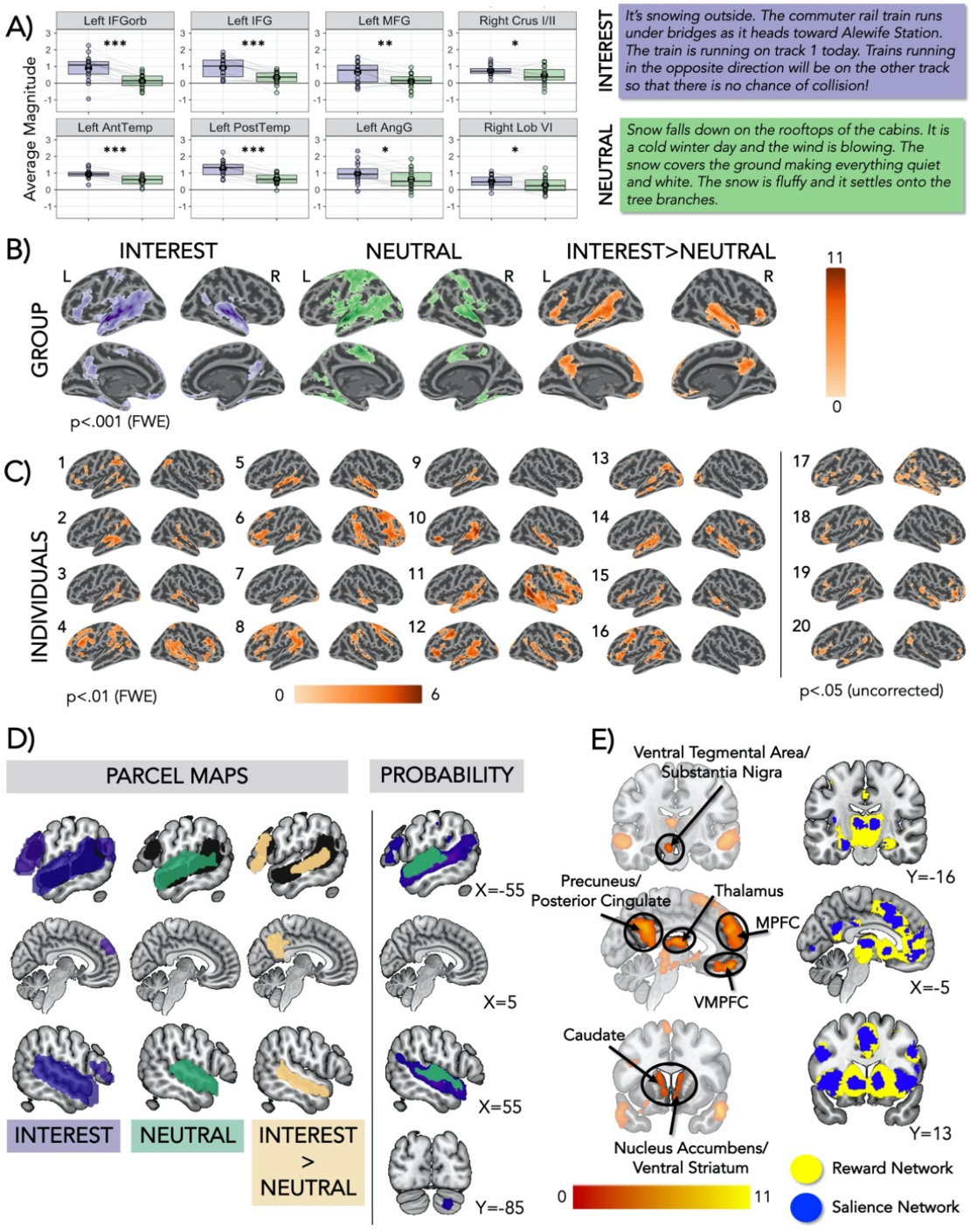
Results

Personally-interesting narratives increased activation in language regions.

To determine whether personally-interesting narratives modulated activation in language regions, we extracted functional responses from *a priori* left frontal, temporal, parietal, and right cerebellar regions of interest (ROIs) canonically associated with language processing (Fedorenko, Hsieh, Nieto-Castañón, Whitfield-Gabrieli, & Kanwisher, 2010). Across language regions, activation was higher for personally-

interesting narratives than for non-personalized "neutral" narratives (main effect of condition: Interest>Neutral: Est.=0.47, S.E.=0.04, t-value=11.69, p<0.001; **Figure 4.1A**).

Figure 4.1: Personally-interesting narratives engage language regions and subcortical regions in neurotypical children.



*(A) Boxplots show average BOLD response to personally-interesting and neutral narratives within 8 language ROIs (right: example narratives). Black circle = mean; gray lines connect individual participants. *p<0.05, **p<0.01, ***p<0.001 (uncorrected).* ***(B)***

Group average for Interest>baseline, Neutral>baseline, and Interest>Neutral (p<0.001, FWE cluster p<0.05). ***(C)*** *Individual whole-brain responses to Interest>Neutral language visualized at p<0.01, FWE cluster p<0.05. Participants who did not show suprathreshold voxels at this threshold or in surface space are visualized at p<0.05 uncorrected.* ***(D) Left****: Parcels within which >80% of participants show significant activation, overlaid on language ROIs (black).* ***Right****: Overlay of probability maps for interest (purple) and neutral (green), each thresholded for 25% overlap at the voxel level.* ***(E) Left****: Group-level activation for Interest>Neutral in classical reward/salience regions.* ***Right****: Neurosynth uniformity maps for "reward" and "salience"; FDR corrected 0.01.*

Given that the personally-interesting narratives feature each child's favorite topic, it was possible that they would indiscriminately increase activation across large swaths of the brain. Instead, a whole-brain analysis revealed that increased cortical activation for personally-interesting narratives was mostly constrained to language regions (e.g., bilateral superior and middle temporal gyri and inferior frontal gyrus), both at the group level (**Figure 4.1B**) and at the level of individual children (**Figure 4.1C**). This result was made all the more striking by the fact that the contrast (Interest > Neutral) presumably controlled for language processing, suggesting that language areas were *specifically sensitive to interest*.

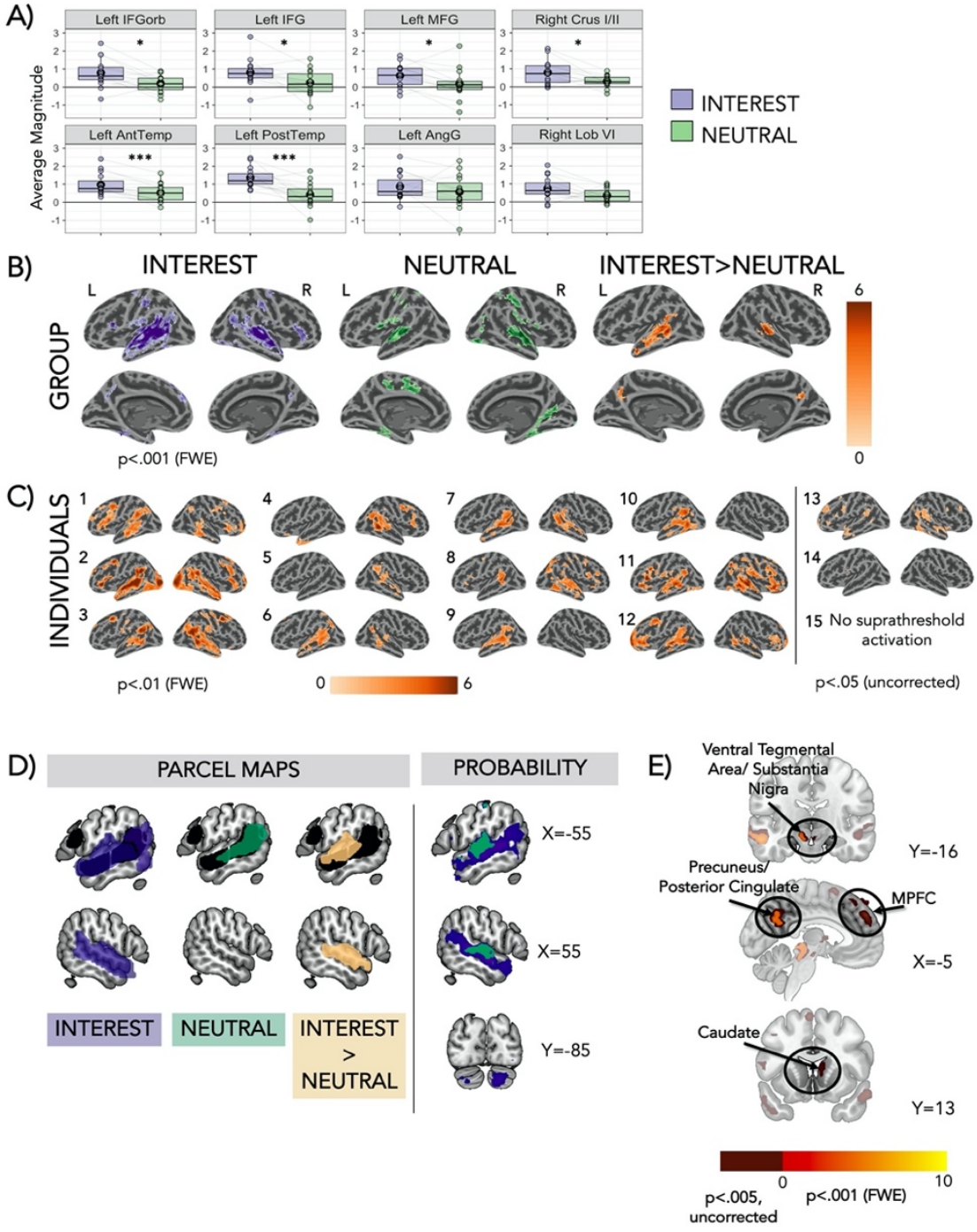
A concern with personalization is that using different stimuli will give rise to discrepant patterns of activation across individuals. Using a data driven approach, we identified large regions (i.e., "parcels") wherein over 80% of subjects showed significant activation. More parcels were identified for personally-interesting than neutral narratives, and these parcels roughly recapitulated canonical language regions, suggesting that idiosyncratic stimuli did not lead to more discrepant activation patterns (**Figure 4.1D, left**). We also examined intersubject overlap at the voxel level, finding more overlapping voxels for personally-interesting than neutral narratives (**Figure 4.1D,**

right). These results suggest that despite the fact that stimuli were idiosyncratic, ranging in topic from train lines to video games, activation patterns for personally-interesting narratives were *more* consistent across participants than activation patterns for neutral narratives. Finally, the whole-brain analysis revealed higher activation for personally-interesting narratives in regions implicated in reward and salience processing, such as the caudate, nucleus accumbens, ventral tegmental area/substantia nigra, ventromedial and medial prefrontal cortex (VMPFC and MPFC, respectively), and precuneus/posterior cingulate (Figure 4.1E). Several of these regions are involved in narrative processing (Silbert, Honey, Simony, Poeppel, & Hasson, 2014), as well as processing of self-referential, autobiographical (Bainbridge & Baker, 2022), or personally relevant materials (Abraham & Cramon, 2009).

Personally-interesting narratives increased activation in language regions in autistic children.

Finally, we investigated whether the potentiating effects of personally-interesting narratives generalized to autism. We scanned 15 autistic children with specific interests and challenges with social communication. As in neurotypical children, personally-interesting narratives elicited higher activation than neutral narratives in language ROIs (main effect of condition: Interest>Neutral: Est.=0.52, S.E.=0.07, t-value=7.27, $p<0.001$; Figure 4.2A) and in the whole brain (Figure 4.2B-C, E). Autistic children also showed more consistent activation patterns for personally-interesting than neutral narratives (Figure 4.2D).

Figure 4.2: Personally-interesting narratives engage language regions and subcortical regions in autistic children.



(A) Average BOLD responses to personally-interesting and neutral narratives within language ROIs. (B) Group averages for each condition. (C) Individual whole-brain responses to Interest>Neutral, visualized as in Figure 1. (D) Left: Data-driven parcels, overlaid on language ROIs (black). Right: Probability maps for interest (purple) and

*neutral (green), thresholded at 25% overlap at the voxel level. **(E)** Group-level activation in classical reward/salience regions.*

Discussion

For the first time, we show that personal interest led to higher activation in children's language regions, as well as select subcortical regions, during passive listening to narratives. The use of personalized spoken passages highlights the power of intrinsically motivating content on the functions of the language network for both neurotypical and autistic children.

Several factors may have contributed to higher responses for personally-interesting narratives in the disparate regions we observed, such as increased attention and arousal, higher intrinsic motivation, and greater personal relevance of and familiarity with the topic. Similar engagement of canonical language regions alongside medial (MPFC, precuneus) and subcortical (e.g., caudate) regions has been associated with processing highly-relevant personalized language stimuli (i.e., greater activation for mothers' voices than unfamiliar voices, Abrams, Mistry, Baker, Padmanabhan, & Menon, 2022), and auditory narrative processing more generally, which may involve similar component processes (Silbert et al., 2014). Another possibility is that the neutral narratives may have elicited lower-than-expected activation due to the context of the task, in which those narratives were interleaved with highly salient, personally-interesting narratives.

A limitation of personalized experiments is that the gain in ecological validity is associated with a loss of stimulus control. In the present study, the content of the personalized narratives differed between participants based on their interests. While it

is not feasible (or necessary) to personalize stimuli in every neuroimaging experiment, it might be an important consideration for 1) populations in which personalization will increase engagement in the paradigm, and 2) studies in which inferences about group or individual differences may be confounded by differing levels of attention to the stimuli materials (e.g., young children, individuals with language or attention disorders). In support of this, some neuroimaging studies in autistic children found that using personalized stimuli (e.g., mother's faces and special interests) led to higher activation in regions that otherwise appeared "underactive" (relative to neurotypical peers) when using non-personalized stimuli (e.g., Foss-Feig et al., 2016; Kohls, Antezana, Mosner, Schultz, & Yerys, 2018; Pierce & Redcay, 2008).

In sum, this study highlights the potential of personally-interesting material to modulate language function in the brains of neurotypical and autistic children, and the feasibility of personalization to evoke consistent brain responses. Future studies might consider personal interest as a powerful tool for maximally probing the scope and functionality of brain networks.

Brief Materials and Methods

All participants (total n=35, n=15 autistic) were screened for the presence of a strong interest and provided links to online videos depicting this interest. Based on these materials, researchers wrote and recorded personalized narratives for each child. In the MRI scanner, all children listened to narratives in three conditions: personally-interesting, neutral, and backwards-language. We compared BOLD activation for personally-interesting and neutral narratives in *a-priori* language regions of interest and across the whole brain, and evaluated intersubject consistency across conditions at the

voxel level and within larger regions. Parents provided informed consent, and children provided assent to participate. This protocol was approved by the MIT Committee on the Use of Humans as Experimental Subjects. Data and materials are available on OSF[21].

Extended Methods

Participants. Data were analyzed from 20 neurotypical children (ages 6.98-12.01 years, mean(SD)=9.35(1.52), 5 female/15 male) and 15 autistic children (ages 8.18 – 13.27 years, mean(SD)= 11.17(1.62), 3 female/11 male/1 nonbinary). All children were native speakers of English, had normal or corrected-to-normal hearing and vision, had no contraindications for MRI (e.g., metal in the body), and had a qualifying special interest (see **Personal interest screening** below). Additional exclusion criteria for the neurotypical children included diagnosis of major neurodevelopmental or psychiatric disorders and language difficulties. In n=9 autistic children scanned prior to the onset of the Covid-19 pandemic, autism diagnosis was confirmed via the Autism Diagnostic Observation Schedule (ADOS) administered by a research-reliable clinician (Lord et al., 2000). Data presented in the current manuscript are a subset of 54 children who were originally recruited for participation (n=27 neurotypical, n=27 autistic). N=19 of the original recruited sample were excluded due to: refusal to participate in the fMRI scan or inability to stay in the scanner past the initial T1 (n=5), excessive motion for the language task (n=12), incidental findings (n=1), and incomplete data (n=1). One participant in the autism group returned post-pandemic since no usable functional data was collected on the first attempt.

[21] https://osf.io/dh3wq/

Personal interest screening. Parents expressed interest in the study via an online screening survey. If a child was potentially eligible (i.e., appropriate age, no exclusions based on the criteria listed above, and parent-reported presence of a significant interest, hobby, passion, or affinity), a member of the research team conducted a phone screening and discussion with parents to (1) confirm eligibility, and (2) ask follow-up questions about the child's interest. Criteria for the presence of a personal interest were as follows: (1) the child must engage with the interest for at least an hour per day on average (or would engage with that interest for the specified amount of time if there were no restrictions in place, e.g., screen time limits), (2) the child must have had the same interest for at least the last two weeks, and (3) the interest had to have associated videos that could be used in the fMRI experiment. Parents, in collaboration with their children, were then asked to provide video clips pertaining to their child's interest, which were then used to create personalized narratives for the fMRI experiment (see **Personalized Stimuli Creation**).

Table 4.1: Participant demographics.

Group	NT	ASD
Age (years)	9.35(1.52) range = 6.98 – 12.01	11.17(1.62) range = 8.18 – 13.27
KBIT Matrices (standard score)	118.65(15.39) range = 87.00 – 140.00	114.93(7.92) range = 104.00 – 133.00
KBIT Verbal Composite (standard score)	121.00(12.99) range = 98.00 – 142.00	107.57(12.70) range = 77.00 – 119.00
SRS Communication (T score)	46.00(6.07) range = 37 – 55	78.93(15.99) range = 51 – 114
Autism Quotient (raw score)	42.47(12.39) range = 19 – 66	95.87(15.20) range = 55 – 124

Age of first word (months)	11.61(3.11 range = 6 – 20.41	14.00(8.32) range = 8 – 42
Age of first sentence (months)	20.41(6.60) range = 12 – 36	23.00(7.78) range = 12 – 42
Age for understanding command (months)	12.18(3.75) range = 6 – 18	14.93(9.64) range = 3 - 36
Interests	• Soccer (n=3) • Baseball and football (n=2) • Basketball • Fishing • Fortnite • Minecraft (n=3) • Pokémon • Lego Marvel superheroes • Animals from *Firefly* PBS show • Baking shows • Musicals • Harry Potter • Art tutorial YouTube channel • Calm paint brushing art videos • Transit Systems	• Soccer • Tennis • Computers • Fortnite • Minecraft • Pokémon • Among Us video game • Lego Ninjago (n=2) • Cartoon voices • Dragons from *How to Train Your Dragon* • Fighting insects • Puppies • Hurricanes/Extreme weather • Trains (local commuter line)

Table shows Mean(Standard Deviation) and range. Age=age at scan, KBIT=Kaufman Brief Intelligence Test (Kaufman & Kaufman, 2004), SRS=Social Responsiveness Scale (Constantino & Gruber, 2005), Autism Quotient (Auyeung, Baron-Cohen, Wheelwright, & Allison, 2008). Age of first word, first sentence, and understanding command were asked via parent report.

Experimental Protocol. Participants completed 1-2 study sessions, which involved behavioral testing and a neuroimaging session. The neuroimaging session included an anatomical scan, a functional run of a task involving watching the participants' selected interest videos and nature videos (not discussed in this paper), a functional run of the personal interest language task (discussed in this paper, see **Personal Interest Narrative**

Task below), and optional additional scans that varied between participants. These options included a resting state scan, neural adaptation tasks involving faces, objects, and auditory words, a separate language task, a diffusion scan, and additional runs of the personal interest tasks. Parents completed a set of questionnaires about their child during the visit including questions about demographic and developmental histories (e.g., language onset), the Autism Quotient (AQ, Auyeung et al., 2008), and the Social Responsiveness Scale (SRS, Constantino & Gruber, 2005). Parents provided informed consent, and children provided assent to participate. This protocol was approved by the MIT Committee on the Use of Humans as Experimental Subjects.

Personalized Stimuli Creation. Parents, in collaboration with their child, provided links to online video clips (e.g., YouTube) that captured their child's personal interest, including timestamps for their child's favorite parts of the videos. We cut seven 16-second clips from the provided videos (capturing each child's favorite part of the videos if provided), and wrote short narratives of the scenes from the selected video clips. A female experimenter (HAO) recorded the descriptions in a sound-proof booth, and the audio files were trimmed to be exactly 16 seconds. Language narratives were approximately matched between participants by avoiding personal pronouns (e.g., "I" or "you"), using simple vocabulary (allowing for interest-specific terms), and using short sentences. Due to the unique nature of personal interests, the personal-interest narratives tended to have more specific nouns — e.g., "Alewife Station" or "Lionel Messi" — than the neutral narratives. Both the personally-interesting and neutral narratives included action verbs and sensorially evocative descriptions. See OSF (https://osf.io/dh3wq/) for the neutral and personally-interesting narrative transcripts for all children with usable data. Total word count, number of words per sentence, number of syllables per word, and number of sentences per narrative were approximately

matched between neutral and personally-interesting conditions (Total word count: M(SD)=39.92(4.21) for personal-interesting across all participants and M=45.14 for the neutral narratives [same across all participants], Number of words per sentence: M(SD)=7.40(1.15) for personally-interesting and M=7.74 for neutral; Number of syllables per word: M(SD)=1.40(.10) for personally-interesting and M=1.23 for neutral, and Number of sentences: M(SD)=5.49(.83) for personally-interesting and M=6.0 for neutral).

Behavioral Measures. Nonverbal cognitive reasoning was assessed via the matrices subtest of the Kaufman Brief Intelligence Test, 2nd edition (KBIT-2, Kaufman & Kaufman, 2004). Language skills were assessed via the verbal composite score of the KBIT-2, including the vocabulary and riddles subtests.

Personal Interest Narrative Task. Participants were asked to passively listen to spoken narratives presented binaurally via MRI-compatible headphones using a block-design paradigm. The task consisted of three conditions: personal interest, neutral, and backwards narratives. In the personal interest condition, participants listened to the personalized narratives about their specific interests. In the neutral condition, participants listened to non-personalized narratives describing nature scenes. Nature content included in the neutral narratives was similarly familiar to all children and unrelated to any child's personal interest. In the backwards condition, participants listened to backwards versions of the neutral narratives in order to account for lower-level auditory features of the narratives. Children listened to 7 narratives (16-seconds each) in each condition. Each narrative was followed by an inter-stimulus rest block of 5 seconds (total of 21 narratives across three conditions and 22 rest blocks). To confirm that children were attending to the task without imposing significant physical or

cognitive demands, we included a low-demand attentional check following each narrative. An image of a panda appeared on the screen directly after the narrative for 1.5 seconds, followed by a blank screen for 0.5 seconds. Children were instructed at the beginning of the study to press a button using their pointer finger via an MRI-compatible button box that they held in their hand every time they saw a picture of a panda. Task order was fixed across participants in the following pattern: personal interest, rest, neutral, rest, backwards, rest, etc. [ABCABC…]. Total task time was 8min 8s.

Acquisition. Data were acquired from a 3-Tesla Siemens Prisma scanner located at the Athinoula A. Martinos Imaging Center at the McGovern Institute at MIT, using a 32-channel head coil. T1-weighted structural images were acquired in 176 interleaved slices with 1.0mm isotropic voxels (MPRAGE; TA=5:08; TR=2530.0ms; FOV=256mm; GRAPPA parallel imaging, acceleration factor of 3). Functional data were acquired with a gradient-echo EPI sequence sensitive to Blood Oxygenation Level Dependent (BOLD) contrast in 3.0mm isotropic voxels in 40 near-axial slices covering the whole brain (EPI factor=70; TR=2500ms; TE=30ms; flip angle=90 degrees; FOV=210mm; TA=7:47).

Preprocessing and Statistical Modeling. fMRI data were preprocessed using fMRIPrep v1.1.1 (Esteban et al., 2019). fMRIPrep is a pipeline developed by the Center for Reproducible Neuroscience that includes motion correction, correction for signal inhomogeneity, skull-stripping, spatial normalization to the Montreal Neurological Institute (MNI)-152 brain atlas, segmentation, and co-registration. Preprocessed images were smoothed in SPM12 at 6mm FWHM. First level modeling was performed using SPM12. Individual regressors for each condition (interest, neutral, backwards, and

button press) were included in the model. Individual TRs were marked as outliers if they had greater than 1mm of framewise displacement. We included one regressor per outlier volume in the first level model, and we excluded participants with > 20% outlier volumes. The critical contrast (interest > neutral) was created to examine regions showing greater activation for personally-interesting than neutral narratives.

Region of Interest Analyses. To determine whether personal interest activated language regions specifically, parameter estimates for each condition were extracted from *a priori* regions of interest (ROIs) known to be important for language processing (Fedorenko et al., 2010). These ROIs are based on an atlas comprised of functional data from 803 participants during language tasks and reflect regions wherein a high proportion of those participants showed overlap in activation patterns (Lipkin et al., 2022). To capture responses in canonical language-selective regions, we selected eight parcels that are commonly associated with language (Fedorenko et al., 2010): left IFGorb, left IFG, left MFG, left AntTemp, left PostTemp, left AngG, right cerebellum lobule VI, and right cerebellum Crus I/II (see below). Linear mixed-effects models were run in R using the lme4 package. To determine if there was an effect of condition (interest, neutral) across the "language network", we used:

$$Y_{BOLDfromROI} \sim X_{condition} + X_{(1|ROI)} + X_{(1|participant)}$$

with participant and ROI as random factors to account for repeated measures. Second, to visualize effects of condition within each language ROI separately, we then used:

$$Y_{BOLDfromROI} \sim X_{condition} + X_{(1|participant)}$$

with participant as a random factor to account for repeated measures.

Group Whole Brain Analysis. Group-level modeling was performed using SPM12. One-sample t-tests were used to determine regions for which activation in each condition of interest (neutral, interest, interest > neutral) was greater than baseline. Group maps were thresholded at an uncorrected voxel $p<0.001$, with a cluster correction for multiple comparisons (FWE<0.05). For comparison, Neurosynth uniformity maps thresholded at FDR<0.01 (Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011) for keywords "reward" and "salience" are presented in Figure 4.1E.

Overlap Analyses. A group-constrained subject specific (GCSS) approach was used to assess consistency and spatial overlap in activation patterns across different conditions (Fedorenko et al., 2010). For each contrast of interest, each participant's statistical parametric map was thresholded voxelwise at $p<0.001$ (uncorrected) and binarized. Binarized maps were overlaid to create a probability map of regions engaged by the contrast of interest, which was then smoothed at 6mm FWHM and thresholded voxelwise at n=2 subjects. Probability maps reflect the number of participants showing overlap in a particular voxel. Secondly, a watershed algorithm from the SPM-SS toolbox was applied to detect local probability maxima from probability maps and extend them spatially to create functionally-defined "parcels". To identify regions within which a large number of participants showed significant activation, we retained parcels which contained significant voxels from 80% or more of participants.

Preregistration. The main hypotheses for the current study were included as part of a broader preregistration in 2018 for a study investigating the neural correlates of personal interest in visual, reward, and language domains in neurotypical and autistic children[22]. Though beyond the scope of the current study, the planned study included additional groups (e.g., a neurotypical group with general but not specific interests), as well as a video task and associated analyses that are not presented here. For the analysis of the personal interest language task, we deviated from the preregistration by not using subject-specific functional ROIs (neutral>backwards), as this would have precluded a comparison between our conditions of interest (personal interest vs. neutral). Instead, we used *a priori* ROIs and whole brain analyses. The following hypotheses were tested and confirmed: 1) All children will show greater activation in the language network for personally-interesting than neutral narratives, and 2) All children will show greater activation in the reward network for personally-interesting than neutral narratives. We did not test hypotheses related to group differences between neurotypical and autistic children, nor associations with behavioral measures, due to smaller than anticipated sample sizes as a result of the COVID-19 pandemic and subsequent data/personnel limitations.

Acknowledgments

This research was supported by the Hock E. Tan and K. Lisa Yang Center for Autism Research at MIT, Seth Klarman and Paul Gannon (to JDEG), NIH F32 MH117933 and Simons Center for the Social Brain Postdoctoral Fellowship (to AMD), NSF Graduate Research Fellowship Program #1745302 (to HAO), and MIT Hugh Hampton Young Memorial Fellowship and MIT Media Lab Learning Innovation Fellowship (to KTJ).

[22] https://osf.io/nr3gk

We thank Hannah Grotzinger for task programming assistance, Caitlin Malloy for ADOS support, and Cindy Li for assistance with recruiting, coordinating, and testing with our autism group. We also thank our undergraduate and high school research assistants who assisted with various aspects of the project, including: Jimmy Chen, Nicole Dundas, Insha Merchant, Rucha Kelkar, Alana Kalehua, and Hillary Jean-Gilles. We are grateful to Atsushi Takahashi and Steve Shannon from the Athinoula A. Martinos Imaging Center at MIT. We thank the participants and families for making this research possible. Finally, we thank Ron Suskind, whose experience with his son Owen inspired this research.

References

Abraham, A., & Cramon, D. Y. von. (2009). Reality = Relevance? Insights from Spontaneous Modulations of the Brain's Default Network when Telling Apart Reality from Fiction. *PLOS ONE*, *4*(3), e4741. https://doi.org/10.1371/journal.pone.0004741

Abrams, D. A., Mistry, P. K., Baker, A. E., Padmanabhan, A., & Menon, V. (2022). A Neurodevelopmental Shift in Reward Circuitry from Mother's to Nonfamilial Voices in Adolescence. *Journal of Neuroscience*, *42*(20), 4164–4173. https://doi.org/10.1523/JNEUROSCI.2018-21.2022

Auyeung, B., Baron-Cohen, S., Wheelwright, S., & Allison, C. (2008). The Autism Spectrum Quotient: Children's Version (AQ-Child). *Journal of Autism and Developmental Disorders*, *38*(7), 1230–1240. https://doi.org/10.1007/s10803-007-0504-z

Bainbridge, W. A., & Baker, C. I. (2022). Multidimensional memory topography in the medial parietal cortex identified from neuroimaging of thousands of daily memory videos. *Nature Communications*, *13*(1), 6508. https://doi.org/10.1038/s41467-022-34075-1

Constantino, J. N., & Gruber, C. P. (2005). *Social Responsiveness Scale*. Los Angeles: Western Psychological Services.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., … Gorgolewski, K. J. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4

Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New Method for fMRI Investigations of Language: Defining ROIs Functionally in Individual Subjects. *Journal of Neurophysiology*, *104*(2), 1177–1194. https://doi.org/10.1152/jn.00032.2010

Foss-Feig, J. H., McGugin, R. W., Gauthier, I., Mash, L. E., Ventola, P., & Cascio, C. J. (2016). A functional neuroimaging study of fusiform response to restricted interests in children and adolescents with autism spectrum disorder. *Journal of Neurodevelopmental Disorders*, *8*(1). https://doi.org/10.1186/s11689-016-9149-6

Harrop, C., Amsbary, J., Towner-Wright, S., Reichow, B., & Boyd, B. A. (2019). That's what I like: The use of circumscribed interests within interventions for individuals with autism spectrum disorder. A systematic review. *Research in Autism Spectrum Disorders*, *57*, 63–86. https://doi.org/10.1016/j.rasd.2018.09.008

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Brief Intelligence Test* (2nd ed.). Circle Pines, MN: American Guidance Service.

Klin, A., Danovitch, J. H., Merz, A. B., & Volkmar, F. R. (2007). Circumscribed Interests in Higher Functioning Individuals With Autism Spectrum Disorders: An Exploratory Study. *Research and Practice for Persons with Severe Disabilities*, *32*(2), 89–100. https://doi.org/10.2511/rpsd.32.2.89

Kohls, G., Antezana, L., Mosner, M. G., Schultz, R. T., & Yerys, B. E. (2018). Altered reward system reactivity for personalized circumscribed interests in autism. *Molecular Autism*, *9*(1). https://doi.org/10.1186/s13229-018-0195-7

Krapp, A., Hidi, S., & Renninger, K. A. (1992). Interest, learning, and development. In K. A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 3–25). Lawrence Erlbaum Associates, Inc.

Lipkin, B., Tuckute, G., Affourtit, J., Small, H., Mineroff, Z., Kean, H., … Fedorenko, E. (2022). Probabilistic atlas for the language network based on precision fMRI data from >800 individuals. *Scientific Data*, *9*(1), 529. https://doi.org/10.1038/s41597-022-01645-3

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., … Rutter, M. (2000). The autism diagnostic observation schedule-generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders*, *30*(3), 205–223.

Pierce, K., & Redcay, E. (2008). Fusiform Function in Children with an Autism Spectrum Disorder Is a Matter of "Who." *Biological Psychiatry*, *64*(7), 552–560. https://doi.org/10.1016/j.biopsych.2008.05.013

Shnayer, S. W. (1968). *Some Relationships between Reading Interest and Reading Comprehension*. Retrieved from https://eric.ed.gov/?id=ED022633

Silbert, L. J., Honey, C. J., Simony, E., Poeppel, D., & Hasson, U. (2014). Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proceedings of the National Academy of Sciences*, *111*(43), E4687–E4696. https://doi.org/10.1073/pnas.1323812111

Sparks, E., Schinkel, M. G., & Moore, C. (2017). Affiliation affects generosity in young children: The roles of minimal group membership and shared interests. *Journal of Experimental Child Psychology*, *159*, 242–262. https://doi.org/10.1016/j.jecp.2017.02.007

Tomova, L., Wang, K. L., Thompson, T., Matthews, G. A., Takahashi, A., Tye, K. M., & Saxe, R. (2020). Acute social isolation evokes midbrain craving responses similar to hunger. *Nature Neuroscience*, *23*(12), 1597–1605. https://doi.org/10.1038/s41593-020-00742-z

Van Lancker, D. (1991). Personal relevance and the human right hemisphere. *Brain and Cognition*, *17*(1), 64–92. https://doi.org/10.1016/0278-2626(91)90067-I

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, *8*(8), 665–670. https://doi.org/10.1038/nmeth.1635

"Not all stories speak to all listeners, but all listeners can find a story that does, somewhere, sometime. In one form or another."

— Erin Morgenstern, ***The Starless Sea***

Chapter 5 : Implementing Remote Developmental Research: A Case Study of an RCT Language Intervention During COVID-19

A version of this chapter has been published as:

[†]Ozernov-Palchik, O., [†]**Olson, H. A.**, Arechiga, X. M., Kentala, H., Solorio-Fielder, J. L., Wang, K. L., Camacho Torres, Y., Gardino, N. D., Dieffenbach, J. R., & Gabrieli, J. D. E. (2022). Implementing Remote Developmental Research: A Case Study of a Randomized Controlled Trial Language Intervention During COVID-19. *Frontiers in Psychology, 12*, 6163. https://doi.org/10.3389/fpsyg.2021.734375
[†]*Authors share joint first authorship.*

Abstract

Intervention studies with developmental samples are difficult to implement, in particular when targeting demographically diverse communities. Online studies have the potential to examine the efficacy of highly scalable interventions aimed at enhancing development, and to address some of the barriers faced by underrepresented communities for participating in developmental research. During the COVID-19 pandemic, we executed a fully remote randomized controlled trial (RCT) language intervention with third and fourth grade students (N = 255[23]; age range 8.19-10.72 years, mean = 9.41, SD = 0.52) from diverse backgrounds across the United States. Using this as a case study, we discuss both challenges and solutions to conducting an intensive online intervention through the various phases of the study, including recruitment, data collection, and fidelity of intervention implementation. We

[23] This was the sample size at the time we wrote this paper. We decided to continue recruiting participants to increase sample sizes when more participants than expected dropped out before posttesting.

provide comprehensive suggestions and takeaways, and conclude by summarizing some important tradeoffs for researchers interested in carrying out such studies.

Introduction

Intervention research in developmental science

One overarching goal of developmental research is to improve children's outcomes. The most direct way to achieve this goal is to implement an intervention - some manipulation of a child's experience or environment - and determine whether it leads to positive changes in outcomes. Not only do such studies allow researchers to test the efficacy of specific intervention programs, but they also play a crucial role in understanding developmental phenomena by elucidating causal mechanisms. A randomized controlled trial (RCT) design is a gold standard for establishing causality and efficacy in intervention research.

Despite the importance of intervention studies in developmental science, executing these studies is difficult. Because effect sizes tend to be small in developmental intervention studies, large samples are needed to detect significant effects (Kraft, 2020; Lortie-Forgues & Inglis, 2019). Interventions must be administered with high fidelity, which can be challenging at a large scale and when they require the involvement of caregivers or educators (Barton & Fettig, 2013; Fixen et al., 2005; O'Donnell, 2008). While in-lab intervention studies allow for highly controlled testing environments, they run the risk of not generalizing to real-world settings (Lortie-Forgues & Inglis, 2019). Additionally, in order to substantially impact a child's experiences or environment, interventions typically have to be implemented over a long period of time (e.g., on the order of weeks to months). Both recruitment and retention of participants in developmental research intervention studies pose significant challenges.

Further, if interventions are to be translated into wide use, they have to be highly scalable to large numbers of children in diverse environments. In particular, the field of developmental research has recently come under scrutiny for predominantly studying WEIRD (western, educated, industrialized, rich, and democratic) populations (Nielsen et al., 2017). Even in the limited context of the United States, participants from lower socioeconomic status (SES) backgrounds are consistently underrepresented in research (Manz et al., 2010; Nicholson et al., 2011), and the majority of developmental science publications do not achieve a race/ethnicity distribution that matches that of the United States population (Bornstein et al., 2013). In addition to the profound issues related to equity (Lorenc et al., 2013; Veinot et al., 2018), lack of diversity and representativeness in developmental science threatens the generalizability of findings and fundamentally hinders our understanding of human development (Nielsen et al., 2017).

One major roadblock to the inclusion of more representative samples is the low participation rates of families from disadvantaged backgrounds in research (Heinrichs et al., 2005). There are multiple barriers to research participation that these families face, including informational barriers (not knowing about research opportunities), perceptual barriers (how families view the purpose and significance of research), and practical barriers such as lack of time and access to transportation (Heinrichs et al., 2005; Whittaker & Cowley, 2012). There are also many hard-to-reach communities in remote areas, far from universities and research centers. Practical barriers are most prohibitive for families from disadvantaged backgrounds (Lingwood et al., 2020).

Online studies: New opportunities for developmental intervention research

Online developmental research studies are becoming increasingly popular and have advanced rapidly during the COVID-19 pandemic. The main benefit of online studies is that they allow families to participate in research from the convenience of their own homes. These studies can take multiple forms, including moderated/synchronous video-based studies (i.e., a live experimenter interacts with a child over a video conferencing platform, such as the Parent and Researcher Collaborative: https://childrenhelpingscience.com; see a review by Sheskin et al., 2020), unmoderated/asynchronous video-based studies (i.e., through platforms that collect video without a live experimenter present, such as Lookit: https://lookit.mit.edu; Scott & Schulz, 2017; for review see Rhodes et al., 2020), and unmoderated app-based studies (Gillen et al., 2021). Despite the increasing popularity of online developmental research and the promise of these methods for increased diversity and scalability (Casler et al., 2013; Kizilcec et al., 2020; Rhodes et al., 2020; Scott et al., 2017), online intervention research is still very limited (but see Kizilcec et al., 2020 for an example).

There are multiple factors to weigh when deciding whether and how to implement an online intervention study. For example, moderated research studies - particularly ones that target underrepresented populations - require a large investment of resources and labor (Rhodes et al., 2020). Using an online platform may increase geographic and racial representation (Rhodes et al., 2020; Scott et al., 2017), but at a potential risk of excluding low-income participants due to a lack of reliable internet and technology (Lourenco & Tasimi, 2020; Van Dijk, 2020). Disparities in access to internet and devices – i.e., the "digital divide" (Van Dijk, 2020) – were particularly apparent early in the pandemic, and concerns were raised about whether online studies would inadvertently decrease diversity in developmental studies (Lourenco & Tasimi, 2020). Finally, implementing research studies in participants' homes, unlike in-lab studies, requires

giving up some control over the study environment. In this paper, we describe some of the important factors to consider in the context of our experience implementing an intensive, fully remote RCT language intervention with third and fourth grade students (ages 8-10 years) from diverse backgrounds across the United States from summer 2020 – spring 2021. Notably, this study used a moderated online study design with extensive direct communication, and thus our suggestions are specific to this particular approach. We conclude by highlighting three main tradeoffs to think about when designing a remote intervention study with a developmental sample.

Case Study: A remote language intervention study during the COVID-19 pandemic

During the COVID-19 pandemic, we implemented an RCT intervention to assess the impact of listening to audiobooks on reading and language skills. Third and fourth grade students were randomly assigned to the Scaffolding, Audiobooks-only, or Mindfulness (active control) group. Children in the Audiobooks-only condition received unlimited access to audiobooks via the Learning Ally platform[24], curated based on their listening comprehension level. Children in the Scaffolding condition also received audiobooks and recommendations, as well as one-on-one online sessions with a learning facilitator twice per week, focused on improving their listening comprehension strategies and supporting their intervention adherence. The Mindfulness group completed a control intervention using a mindfulness app. The intervention period was 8 weeks for each group, with 2-3 hours of pre-testing and 2-3 hours of post-testing using a battery of measures administered via Zoom. We believe that this project will serve as an informative case study for other developmental researchers considering adapting intensive developmental interventions to an online format. Hypotheses,

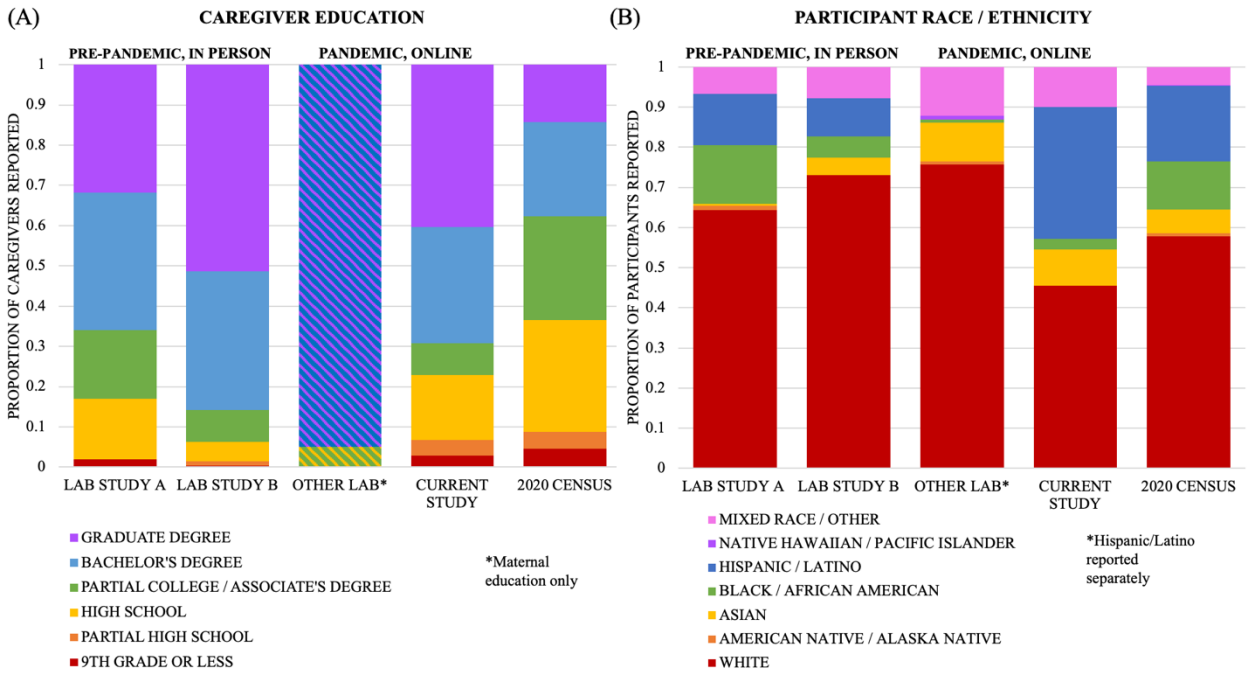
[24] https://learningally.org/

detailed methods, and results from the study will be presented in a separate manuscript (Olson, Ozernov-Palchik, et al., manuscript in preparation).

Recruitment

An important consideration for developmental researchers planning an online intervention study is whether they will be able to recruit a large enough sample size within a feasible time frame. Furthermore, researchers may be looking to recruit samples that are representative in terms of demographic variables like race/ethnicity and SES. As a case study, we will first describe our final sample characteristics, and then outline specific examples of recruitment efforts throughout the study period that led to this sample, including costs for various recruitment strategies.

Figure 5.1: Demographic Comparison to Three Representative Studies.



Summary information describing two studies from our lab conducted prior to the pandemic (Lab Study A, Ozernov-Palchik et al., 2017; Lab Study B, Pollack et al., 2021), and one from another lab conducting a similar study during the pandemic (Other Lab; Bambha & Casasola, 2021). For Lab Study B, we included all participants who completed any portion of the study. ***(A)*** *Highest level of parental education attainment, including both parents, for all who responded (Lab Study A, N=358; Lab Study B, N=463; Other Lab [maternal only], N=118; Current Study, N=449). 2020 Census includes all adults 25 years and older.* ***(B)*** *Parent-reported race/ethnicity of the child, for all who responded (Lab Study A, N=179; Lab Study B, N=230; Other Lab, N=115; Current Study, N=231). Participants who identify as Hispanic/Latino are counted in that category, regardless of race. Other categories reflect that race alone (not Hispanic/Latino).*
Note: Bambha & Casasola reported maternal education only: obtained high school degree (118/118), obtained 4-year college degree or above (112/118); and reported Hispanic/Latino separately from race (15/115 were Hispanic or Latino).

Table 5.1: Comparison to Three Representative Studies.

Study	N	Age Range	Setting	Recruitment	Time	Type
Lab Study A	182	8-10 years	Lab	School Partnership	Pre-pandemic	Neuroimaging/ Longitudinal
Lab Study B	248	8-13 years	Lab	School Outreach + Social Media	Pre-pandemic	Neuroimaging
Other Lab	118	3-5 years	Online	Social Media	Pandemic	Intervention
Current Study	255	8-10 years	Online	School Outreach + Social Media	Pandemic	Intervention

Participants

Beginning in mid-summer 2020, we set out to recruit 240 third and fourth grade students (80 per group) with a broad range of demographic, geographic, reading level, and SES characteristics. To be eligible for the first pre-testing session, children had to be fluent in English, have a caregiver who spoke English or Spanish, and have no diagnosis of autism spectrum disorders or hearing impairments. Given that all sessions were held virtually, over Zoom, we unfortunately could not accommodate families who

did not have internet or computer/tablet access (*N* =14). However, because this study took place during the pandemic, many school systems provided children with access to these resources. We reached back out to families who expressed interest but initially lacked a computer and/or internet over the summer to see if they had been provided these resources by the school system during the school year. Since many families in poor and rural communities lack access to reliable internet (Lourenco & Tasimi, 2020; Van Dijk, 2020), our sample may not be representative of the most severely-affected lower-income communities. Children were compensated $20 per hour for all pre-testing and post-testing sessions (approximately six hours total during the study). Caregivers were additionally compensated $5 per survey for completing a total of ten surveys at the beginning and end of the study. Families also received lifetime access to the Learning Ally audiobook service after completion of the study, regardless of their group assignment.

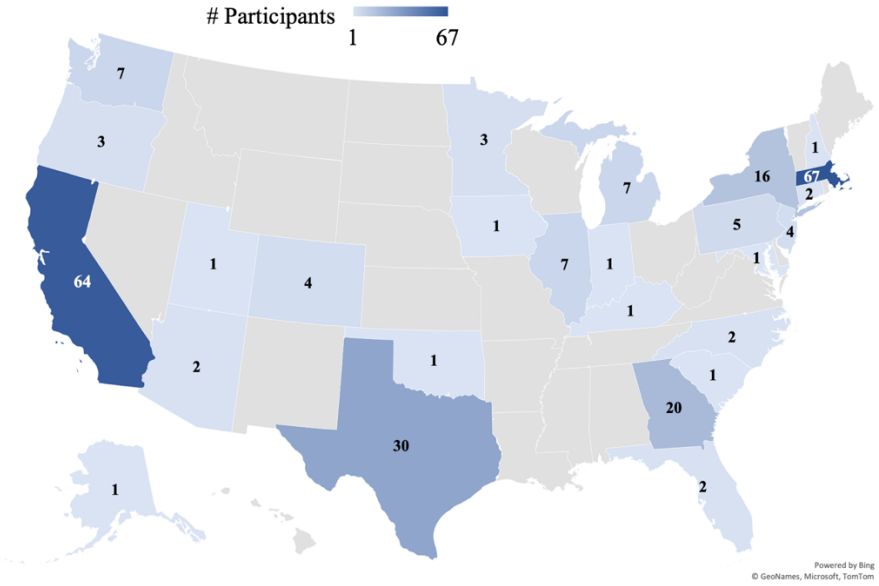
Figure 5.1 shows demographic information for the 255 participants (age range 8.19-10.72 years, mean = 9.41, SD = 0.52) who were eligible for our study and were included in one of our three intervention groups, as well as how our sample compares to the US Census data from 2020 (excludes participants who did not respond to these questions; NA=24 for race/ethnicity, NA=24 for Parent 1 education, NA=37 for Parent 2 education). To demonstrate how the sample demographics in this study compare to similar in-lab and online studies, we also show demographic distributions from three comparison studies (Table 5.1; Figure 5.1): a pre-pandemic longitudinal neuroimaging study conducted in our lab that relied on school partnerships and in-school testing for recruitment prior to the pandemic (Lab Study A, Ozernov-Palchik et al., 2017), a neuroimaging study conducted in our lab that used a combination of outreach events, advertisements, and social media to recruit participants (Lab Study B, Pollack et al.,

2021), and an online intervention study conducted by another lab during the pandemic (Other Lab, Bambha & Casasola, 2021). We conducted a chi-square analysis to compare differences in the frequency of children with parental education of only high school between the current study and the four comparison samples (i.e., Lab Study A, Lab Study B, Other Lab, Census). The current study was not significantly different in the frequency of high school level education or below than the Lab Study A ($X^2(1) = 3.12$, $p = 0.078$) and Lab Study B ($X^2(1) = 0.3$, $p = 0.584$), but it had higher frequency of high school level education or below than the Other Lab study ($X^2(1) = 26.15$, $p < 0.001$) and lower frequency than the 2020 US Census data ($X^2(1) = 76.6$, $p < 0.001$). For a study conducted entirely online and during the pandemic, we successfully achieved a socioeconomically diverse sample comparable to pre-pandemic in-person studies that relied on in-school recruitment. Notably, the comparison online study – which did not specifically aim to recruit a diverse sample in terms of SES – included almost all mothers with at least a 4-year college degree. Thus, the transition to online studies does not automatically increase participant diversity in terms of SES.

We also evaluated differences in the frequency of white participants across the five samples. Our study had a lower frequency of white participants than Lab Study A ($X^2(1) = 13.58$, $p < 0.001$), Lab Study B ($X^2(1) = 35.19$, $p < 0.001$), the Other Lab study ($X^2(1) = 27.14$, $p < 0.001$), and the 2020 US Census ($X^2(1) = 14.02$, $p < 0.001$). The majority of developmental studies do not have representative samples in terms of racial diversity (Bornstein, Jager, & Putnik 2013). There are important caveats to the comparison between the current study and the other lab studies, however. The in-lab studies were not conducted during a pandemic, and they involved neuroimaging. Despite their longitudinal nature, the in-lab studies did not include an intervention, which may have incentivized participation from some families. Nevertheless, although the comparison is

not well-controlled, it suggests that we were successful in recruiting a diverse, representative sample of participants. Furthermore, we attained substantially more geographic diversity than is possible with in-lab studies. Our 255 participants came from a total of 26 states and 186 zip codes in the United States, plus Canada (**Figure 5.2**).

Figure 5.2: Map of Participants by State.



Map shows number of participants per state that were sorted into one of the three intervention groups (N=255). Not shown: 1 participant from Canada.

Overall Recruitment Strategies

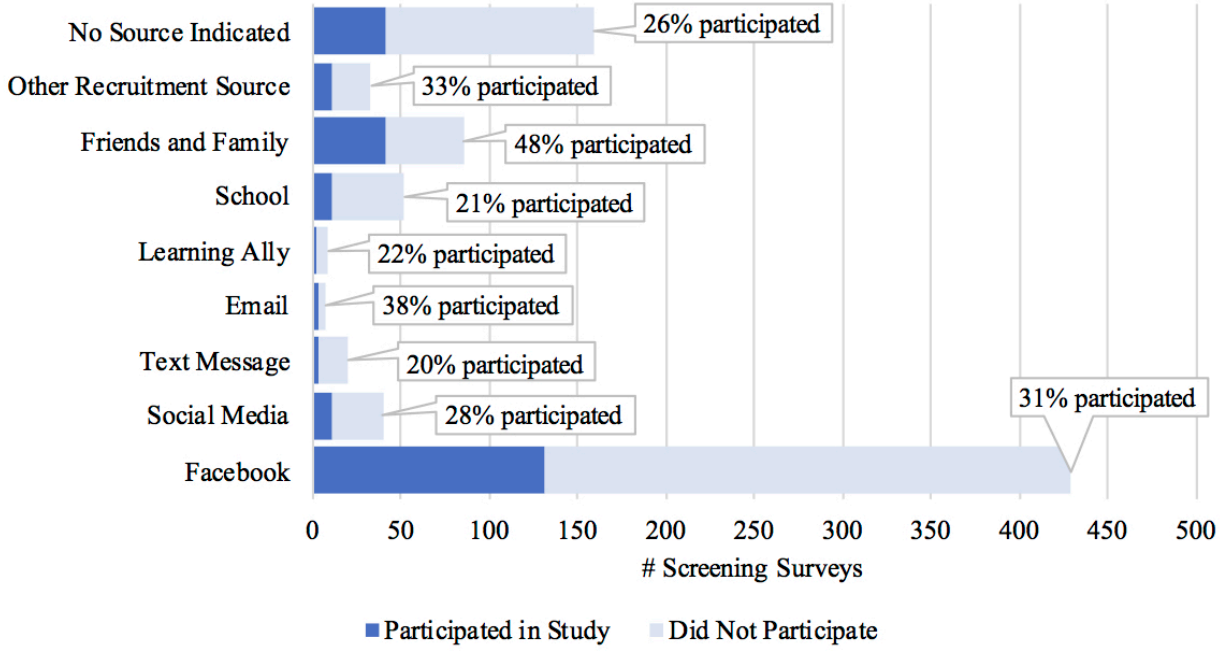
To attain a diverse sample for our online intervention study, we tried several avenues for recruitment, including existing relationships with schools, new school partnerships, and online advertising. We received MIT Institutional Review Board (IRB) approval for all of our recruitment materials including flyers and social media ads in English and Spanish. These flyers and ads included a link directing caregivers to our participant

screening survey. All study data, including data from the screening survey, were managed using REDCap (Research Electronic Data Capture), a secure, web-based software platform designed to support data capture for research studies (Harris et al., 2009, 2019). The landing page, available in English and Spanish, briefly outlined the study and asked the parent or guardian to provide contact information, a simple demographic profile of their child, and other factors relevant to the study (e.g., access to technology). We included the question, "Does your child receive free or reduced lunch at school?" and prioritized contacting the families that responded 'yes' to this question. Below, we describe the efficacy of our different recruitment strategies, as well as our takeaways for other researchers considering these methods for an online intervention study.

School Partnerships. We began recruitment efforts in summer 2020 by reaching out to large and diverse school districts with whom we had existing relationships. Our hope had been to disproportionately recruit lower SES students based on the profiles of the districts, such as public schools with high percentages of free/reduced lunch eligibility. We met with district leaders and principals, who expressed their enthusiasm and commitment to supporting our study. Fourteen schools, all with a large proportion of free/reduced lunch eligible families, officially partnered with our study. Outreach efforts by educators at our partner schools included pre-recorded phone calls to families, flyers, and text messages, with a range of 3-8 outreach attempts per school to their eligible students. This outreach yielded a relatively small fraction of the target number of students (Figure 5.3). It is important to note, however, that our school recruitment efforts took place during the early months of the pandemic when many educators were managing the logistics of school closures, and caregivers were getting accustomed to the new realities of remote learning. Additionally, our school partnership efforts were

limited to schools with predominantly English and Spanish speaking parents and caregivers, as we were not able to accommodate families in additional languages. Online intervention studies that choose to focus their recruitment on specific school districts should likewise consider the predominant language(s) spoken within the community, as we found that our study required substantial ongoing communication with families to provide appropriate support and ensure adherence (see Family Communication and Retention, below).

Figure 5.3: Completed Screening Surveys and Final Participants by Recruiting Source.



Social Media. Our biggest recruitment success came from social media advertising through Facebook and Twitter. However, recruiting via these modalities introduced a unique set of challenges and considerations. One other online option we pursued was Craigslist targeted for specific zip codes, but this approach was ineffective due to Craigslist's stringent policies regarding the categorization of ads.

Facebook. We first posted about our study on our lab's Facebook page. Our lab had existing relationships with parent advocacy groups and other organizations that serve students with language-based learning disabilities. These organizations were more likely to include families from higher-SES backgrounds, so our initial social media recruitment efforts were skewed towards this demographic. We then transitioned to paid Facebook ads. Our initial push was not as fruitful, primarily due to a low budget: we originally invested $25 per posted ad, with each post spanning 3-5 consecutive days within a week. Each week, we launched a different ad until we exhausted our 3 differently-themed ads (each available in English and Spanish), then started the sequence over again. After a month, we increased the budget to $300 per posted ad for subsequent weeks. With this latter approach, we settled on 3 consecutive 24-hour days, usually Friday-Monday. Table 5.2 summarizes Facebook ad effectiveness for different representative configurations of ads.

Not surprisingly, it quickly became apparent that the amount of money invested resulted in increased study interest; the higher the investment, the more the ad is advertised across Facebook, Instagram, and Facebook messenger. The more the post is advertised, the greater the opportunity for engagement, and ultimately increased participation numbers. For future studies, if using Facebook, we recommend a generous social media budget to yield a large pool of participants. In total, we spent $4,389 on Facebook advertisements over the course of the study, and a total of 131 of our 255 participants indicated that they found out about our study via Facebook (Figure 5.3), resulting in an average cost of approximately $34 per participant recruited via Facebook (Table 5.2). However, the actual cost per Facebook-recruited participant varied widely during different ad campaigns (Table 5.2).

To help us recruit participants from lower-SES backgrounds, we used targeted advertising. Facebook provides an option to target specific audiences by selecting cities, zip codes, educational level, age of child, individual interests, and more. While more individuals from targeted communities will see the post across their social media accounts, it does not necessarily mean that each individual who engages with the post will enroll in the study, so consistently posting is key to increasing enrollment rates. For instance, after boosting our recruitment success by targeting ads at 25-mile radius circles around select cities (variously, Atlanta, Boston, Chicago, Dallas, Detroit, Houston, Los Angeles, Miami, New York, Philadelphia, Phoenix, and San Antonio), to target families closer to urban centers, we narrowed the radius to 10 miles in an attempt to recruit more lower-SES participants. Recruiting to this profile proved less successful than it was for the 25-mile radius group. We then used a "household income by zip code" list to try to further improve lower-SES recruitment, but as with the 10-mile radius effort, this approach was not successful. Table 5.2 shows estimated costs per participant (qualified and began the intervention) who learned about our study from one of the three ad campaigns. It should be noted that these estimates rely on open-ended report of how participants learned about the study, and that these ad campaigns proceeded sequentially over different times during the year, with substantial variation in exactly which areas were targeted. Thus, while we think the estimates are informative for researchers considering these strategies, many factors likely influenced the number of participants we recruited.

Table 5.2: Effectiveness for Three Representative Facebook Ad Configurations.

Ad Configuration	Total Spend	Impressions	Clicks	Clicks per Thousand Impressions	Cost per Click	Cost per Participant
Set A: 25-mile radius around select cities						
English Ads	$1,714	273,448	3,030	11.1	$0.57	n/a
Spanish Ads	$363	78,593	709	9.0	$0.51	n/a
English + Spanish Ads	$2,077	352,041	3,739	10.6	$0.56	$17.02
Set B: 10-mile radius around select cities						
English Ads	$1,089	160,038	1,823	11.4	$0.60	n/a
Spanish Ads	$373	61,952	524	8.5	$0.71	n/a
English + Spanish Ads	$1,463	221,990	2,347	10.6	$0.62	$86.05
Set C: low SES zip codes						
English Ads	$579	99,180	579	5.8	$1.00	n/a
Spanish Ads	$271	37,984	212	5.6	$1.28	n/a
English + Spanish Ads	$849	137,164	791	5.8	$1.07	$283.06
TOTAL	$4,389	711,195	6,877	9.7	$0.64	$33.50

Total spend, number of advertisement impressions, number of clicks on our screening survey, number of clicks on our screening survey per thousand ad impressions, and the cost per click on our screening survey are shown for three of our Facebook advertisement campaigns. Estimated cost per participant was calculated based on participant report of how they found out about our study on the screening survey (N=255 total participants began the intervention).

Twitter. Learning Ally, the non-profit audiobook company that we partnered with for the study, advertised our study via Twitter (**Table 5.3**). We attribute their much higher ad engagement (about 10x what we saw with Facebook) to their large and strong following. This higher ad engagement did not translate to more sign-ups, however, as no participants explicitly identified Twitter as how they found out about our study.

Table 5.3: Effectiveness for Twitter Ads.

Ad Configuration	Total Spend	Impressions	Clicks	Clicks per Thousand Impressions	Cost per Click
Total campaign	~$450	~20,000	1,793	91.0	$0.25

Takeaways

School partnerships allow for greater control over participant demographics, as researchers can choose to partner with schools that have specific demographic profiles. However, establishing these partnerships takes time and effort, and may yield modest recruitment for an intensive, out-of-school intervention program. While it is certainly possible to establish school partnerships for an online intervention study, it does require substantial resources (both time and money) from the research team. Social media advertising brings the benefits of both large reach and precision targeting. Since online intervention studies do not have geographic constraints, this recruitment strategy may be beneficial for other developmental researchers considering implementing an online intervention.

Response rates per ad shown are quite small – close to 800,000 people viewing the ad yielded less than 150 actual participants. For the paid advertising, the cost ranged from $0.25 to $1.40 per ad click. This relatively wide difference reflects whether the audience knows the advertiser (in the case of Learning Ally's Twitter audience), how the ads were targeted by SES level (lower SES clicks had a higher cost), and what language the ads were in (English had a lower cost than Spanish). It is important to note that

clicks do not remotely equate directly to study participants – the vast majority of people reaching the screener landing page (95%+) did not sign up for the study.

Overall, our recruitment efforts led to a representative sample of participants in terms of caregiver education and child's race/ethnicity (**Figure 5.1**). We also attained substantial geographic diversity, with participants from 186 different zip codes and 26 different states in the United States, plus Canada (**Figure 5.2**). Our sample was not substantially more diverse in terms of caregiver education compared to other studies run by our lab that aimed to recruit diverse samples, but it was more diverse than another similar study run during the pandemic that did not explicitly aim to recruit a diverse sample based on caregiver education. Our sample was also more ethnically/racially diverse than similar in-lab studies and the general United States population. Thus, the transition to an online intervention format does not necessarily lead to more diverse samples on all dimensions without explicit efforts on those fronts, as well as a considerable recruitment budget.

Family Communication and Retention

Another factor developmental researchers will need to consider when adapting to an online protocol for intervention studies is how to ensure continued engagement and adherence to the program. During our study, not only were we collecting data and administering an intervention online, but we were also doing so during a global pandemic. Families dealt with illness, death, financial stress, technological challenges, and other difficulties over the course of the study. We adapted our communication protocols to be as supportive to families as possible. We believe that these lessons are

also worth sharing, as even in non-pandemic times, families encounter these and other challenges.

Personalized communication methods

Having robust procedures for family scheduling and communication was vital to our study. We had a dedicated research team whose primary role was to contact families and answer any questions that came up. This team included two full-time research staff, as well as 2-3 undergraduate research assistants available to troubleshoot specific questions regarding the use of the audiobook app. We received 15-35 emails per day regarding scheduling, rescheduling, payment requests, score report updates, app issues, etc.

Before the study began, we drafted email and text templates for key communication points at various stages before, during, and after the intervention. For example, we had templates for program orientation and onboarding procedures, appointment confirmation and session reminders, as well as periodic check-ins. In our screening form, we asked for each family's preferred method of communication, and we used this method throughout the study. To ensure consistency in communication, one researcher was assigned to each family and handled all communication for that family. While communicating with families using various methods (i.e., emails, phone calls, text messages) was more time and labor intensive, we found that it boosted participation throughout the duration of our study. We observed high retention rates overall, but there was still attrition (**Figure 5.4)**. Text and email reminders helped minimize missed appointments. If participants missed a session or were generally more challenging to

communicate with, we noted this for their next session and asked the tester to send an additional reminder the day of the testing session to ensure attendance.

For the Scaffolding Group in our study (i.e., the intervention group that met biweekly with 'learning facilitators' in addition to listening to audiobooks), the average family required approximately 37 points of contact throughout the study. This included appointment confirmations, reminders about reading books, payment details, and parent surveys. Similar levels of communication were required for the other groups (i.e., Audiobook-only and Mindfulness), with around 24 points of contact per family. Importantly, however, the number of contact points per family within each group varied based on families' circumstances. Families with limited access to and knowledge of technology at home required additional support throughout the study from our research team. Families with more variable work schedules were more likely to miss sessions or need to reschedule. Thus, we strongly advocate for clear, consistent, and individualized communication with all families, which may especially affect the enrollment and retention of the participants from disadvantaged backgrounds.

Importance of bilingual research personnel. There was a large proportion of Spanish-speaking families in our partner schools. Our final sample included 12% Spanish-speaking participants (30/255), and we had two bilingual Spanish-speaking full-time researchers to support these families. At the beginning of the study, there was a large effort to translate all study materials, surveys, and additional resources into Spanish. Although most of the translation effort was front-loaded, there was still a need for Spanish-speaking researchers throughout the study for family communication.

Scheduling. Since we had families from across the United States participate in our study, we had to account for multiple time zones when scheduling sessions with testers and learning facilitators. We were able to schedule sessions around each family's schedule, including weekend and evening sessions. Each tester had a personal, secure Zoom link that was sent to the family before their scheduled session. Unlike in-person data collection, there was no limit to how many sessions we could book at one time, since physical space was not an issue. Testers called and attempted to troubleshoot with the family if the participant had difficulty getting onto Zoom. The child could complete sessions on a computer or tablet; we also allowed children to log onto the Zoom session via a cell phone in circumstances where no other option was available (only for tests without visual stimuli, as image size would be significantly reduced on a phone screen).

Retention

Most families who expressed initial interest by filling out our screening survey did not end up participating in our study. We experienced high attrition between screening, pre-testing, and group assignment. However, once participants completed onboarding procedures and began the 8-week intervention, attrition was quite low (Figure 5.4).

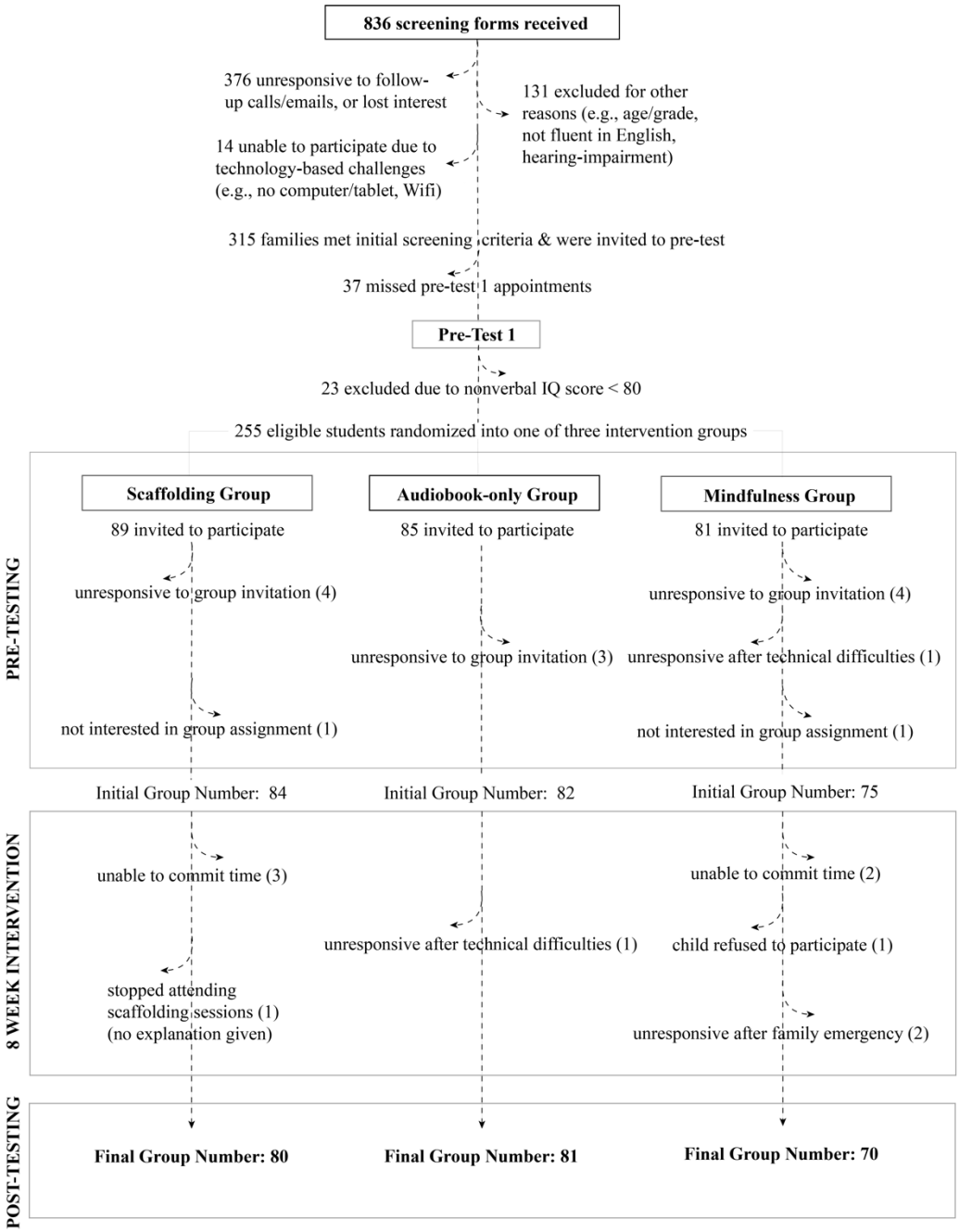
Flexible accommodations to families' individual needs minimized mid-study attrition, but it could not be entirely prevented. In some cases, children were very resistant to participating in the intervention. For example, given the new distance-learning protocols implemented during the COVID-19 pandemic, some children reported not wishing to have more screen-time. This may be relevant for future studies if educators continue to rely on screen-based technology for learning in and out of school. In

instances where children were especially resistant to participating, we did not pressure them to continue. In other instances, however, families simply stopped responding to emails and texts. We observed the greatest non-responsiveness at the end of the program, when attempting to schedule post-test sessions. When faced with non-responsive participants, we first followed-up with multiple (~5) reminder emails, phone calls, and/or text messages, then we issued one final check-in email before suspending any further attempts to reach out.

Takeaways

Overall, we credit the efficiency of our communication pipeline to the use of pre-drafted email/text templates and maintaining an active log of all communications. We recommend frequent and consistent communication with participants to minimize attrition when conducting large-scale online intervention studies. Being timely with responses encourages participants to continue with the study and increases their participation at the post-testing portion of the study. To manage a large number of participant questions, it is important to have a main contact person for each family. We found that regular interaction with our families, via their preferred mode of communication, was effective in establishing rapport and maximizing engagement. Moreover, it is critically important to have bilingual staff who can closely support families who may speak another language. Finally, detailed study orientation materials and clear, step-by-step onboarding procedures are useful to ensure that participants understand all study requirements and to preemptively troubleshoot potential barriers to participation.

Figure 5.4: Participant Pipeline and Attrition.



Data Collection

For developmental researchers that typically utilize in-lab assessments, a major adjustment when transitioning to online intervention studies is adapting measures for

online administration. Here, we describe the measures we used, how assessment scores compared to in-lab administration of the same assessments, how we dealt with variable testing environments, and how we trained our team to administer assessments online.

Behavioral Battery Adaptation

Adapting assessments for online administration required careful consideration to ensure feasibility for both testers and participants. We decided to administer all assessments over Zoom, which allowed testers to directly interact with participants in real-time. For scoring purposes, we audio- and video-recorded each session and stored these recordings securely. The Zoom platform enabled testers to share their screens, allowing us to display scans of stimulus items and online assessment platforms.

Online administration of the assessments in our battery required various considerations and adaptations (Table 5.4). Some tests had already been adapted for online administration, and we used the publisher's online administration and scoring platform. Other tests required tracking the child's responses and simultaneous scoring that was not viable via the computer. We mailed packets to each tester containing printouts of these assessment score sheets along with dry-erase markers and plastic protector sheets. This packet also included a copy of the testing manual containing the required materials and procedures for all tests. These materials allowed testers to have fewer files open on their computer at once during the session. We used DropBox (MIT provides large storage space to its affiliates) to upload all materials for tester access (e.g., stimulus item scans, administration guidelines, etc.), and we used team Slack as a way to troubleshoot or to ask questions before, during, or after test administration.

Table 5.4: Assessments and Adaptations for Remote Administration.

Assessment	Description	Adaptations	Sample Reliability Coefficients	Publisher Reliability Coefficients
Kaufman Brief Intelligence Test, 2nd Edition (KBIT-2) – Matrices[1]	Standardized nonverbal IQ assessment	Scan of stimulus items screen-shared via Zoom.	α: 0.83 split-half: 0.81	split-half: 0.81-0.88
Clinical Evaluation of Language Fundamentals, 5th Edition (CELF-5) - Understanding Spoken Paragraphs[2]	Standardized test of listening comprehension	Administered via Zoom.	α:: 0.74 split-half: 0.79	α:: 0.75-0.85
Dynamic Indicators of Basic Early Literacy Skills (DIBELS)[3] ● Word Reading Fluency (WRF) ● Passage Reading Fluency (PRF) ● Multiple Choice Reading Comprehension (MCRC)	Standardized measures to assess reading skills; MCRC is a computer-administered standardized test	WRF & PRF: Digital forms screen-shared via Zoom. Tester recorded errors on online progress monitoring site from publisher. MCRC: Tester screen-shared and child was given control of tester's screen to select multiple choice answers. Alternative was to have child orally tell tester which answer to select (when child was unable to utilize "Remote Control").	Item level data was not available	

Peabody Picture Vocabulary Test, 5th Edition (PPVT-5)[4]	Standardized receptive vocabulary assessment	Images screen-shared via Zoom using publisher materials adapted for digital use (Q Global).	α:: 0.96 split-half: 0.96	α:: 0.97
Wechsler Abbreviated Scale of Intelligence, 2nd Edition (WASI-II) – Vocabulary[5]	Standardized vocabulary assessment	Scan of stimulus items screen-shared via Zoom.	α:: 0.8 split-half: 0.82	split-half: 0.88-0.93
Comprehensive Test of Phonological Processing, 2nd Edition (CTOPP-2) - Nonword Repetition, Memory for Digits, Blending Words[6]	Standardized measures to assess baseline working memory skills	Audio files sent to families to download ahead of time; child/caregiver asked to play each file from their computer during assessment.	NWR α:: 0.73 split-half: 0.76 MD α:: 0.8 split-half: 0.84 BW α:: 0.84 split-half: 0.86	α:: 0.77 α:: 0.8 α:: 0.8

1. Kaufman, 2004; 2. Wiig et al., 2013; 3. Good et al., 2002; 4. Dunn & Dunn, 2007; 5. Wechsler, 2011; 6. Wagner et al., 1999)

Note: α: represents the Cronbach's alpha and split-half represents the Spearman-Brown prophecy formula. Reliability coefficient values above 0.71 are considered acceptable (George & Mallery, 2003). Publisher reliability information was obtained from the technical manuals and reports released by the respective companies.

Tester Training

The testing team consisted of graduate students from speech and language pathology or early education programs with experience administering psychoeducational evaluations to school-aged children. All testers were native English speakers, and some were also fluent in Spanish. All testers had prior knowledge of the Zoom platform and different file storing/sharing programs (e.g., DropBox, Google Drive). Testers were trained remotely on administering and scoring our assessment battery. Before starting their first session, testers scored a video-recorded session and were deemed ready if they achieved 95% reliability with the first scorer (an experienced tester). A team member reviewed and scored the video recording of each tester's first session with a child and gave them feedback as necessary. Training continued until the testers were able to administer and score all assessments with high accuracy. Testers were blind to participants' group assignments. One benefit of online testing is the ability to easily video record testing sessions. Doing so helped facilitate a more thorough reliability assurance than for in-lab studies that tend to only audio-record sessions.

Remote Administration

We also needed to adapt our general assessment administration procedures. Each session began with the tester confirming the child was in an optimal testing environment, and adjustments were made if necessary (i.e., moving to a quieter space in the home). Caregivers were asked for their permission to have the Zoom session recorded. The tester then reviewed the consent form with the caregiver and the assent form with the child, which had been emailed to the family before the session, and obtained verbal consent from both the caregiver and child. If the family's primary

language was Spanish, the initial session was scheduled with a bilingual tester, or another bilingual member of the team joined the session to obtain consent in Spanish. Testers then administered the assessments. These were split across 2-3 sessions, as the battery of tests was extensive, and children generally fatigued after about 90 minutes. Immediately following the session, testers uploaded the recordings of both the verbal consent/assent and the testing session to a secure server, and submitted records of participants' responses.

Finally, we needed to establish data management and scoring procedures that ensured accuracy in the online setting. Since paper record forms could not be centrally stored with all of our testers working remotely, we created Google Forms to record participants' responses for most assessments. Having digital copies of item-level responses helped with easily calculating reliability for each assessment (Table 5.4). The Google Forms were used to generate spreadsheets of participant data for each assessment. All records only used participant IDs. Other assessments required the use of the developer's platform for scoring.

Testing Environment. The testing team encountered a variety of challenges unique to the virtual testing environment. In-person assessment allows for more knowledge of and control over what participants are doing during the session. With online administration, we relied more on children and caregivers to achieve consistency in the testing environment. For instance, during the online sessions, we needed to make sure that participants could see and hear what we expected them to, despite not having direct control over the visual display and audio output of their devices. Thus, testers regularly asked participants to confirm that they could see the screen-shared materials

and hear their voice clearly, adjusting the size of materials on display and asking children to adjust their speaker/headphone volume as necessary.

The most common issues were loud background noise in the home and poor internet connection, which often affected audio quality for the participant, tester, or both. It was sometimes difficult to judge the quality of what the child was hearing, especially when caregivers were not present to provide feedback. For assessments that involved timed performance or stimulus items that could not be repeated, testers made adjustments to reduce validity concerns. If there was background noise and the child did not have a quieter space, testers asked the child to put on headphones or saved listening tasks for the following session when the child might be in a quieter environment. Child responses were often difficult to discern when answer choices involved rhyming letters (e.g., A, B, C, D), even after asking the child to repeat the response. In these instances, testers requested that the child type their answers into the chat on Zoom.

Internet connectivity and other technical factors (e.g., the ability to download and play audio files provided by the team) varied widely across participants and between sessions. Sometimes testers turned off the video portion of the Zoom call in an attempt to improve the audio connection. The team also encountered minor technical issues with specific aspects of the online administration process, such as problems with using the "Remote Control" function on Zoom on certain types of computers.

At times, the participant's home environment was distracting for other reasons, such as family members or pets entering the room. Many children completed the testing from a desk, but many others completed it while sitting on a couch or in their bed, and some children needed reminders to sit up or change position to better focus on assessment

tasks. Because some caregivers chose to remain in the room during testing, testers occasionally encountered caregivers who continued to help their child despite the tester's requests not to. In particular, because caregivers were often off-camera, it was sometimes difficult to gauge the extent of the support given by the caregiver. The presence of caregivers in the room may have made some children more self-conscious about their performance, whereas other children appeared comforted by their presence. Also, because the tester could not see the child's screen, some children may have attempted to look up answers to certain testing questions, though we do not believe this to be a significant issue overall. The ability to record and re-watch sessions while scoring was critical given these challenges unique to the home setting.

Finally, some children felt fatigued during sessions scheduled after the child had just spent several hours on the computer during remote learning. Testers offered breaks and/or ended the session based on their judgement of the child's fatigue and engagement.

Scoring. To ensure validity, each assessment was double-scored by another tester. The second scorer watched session recordings (stored and accessed on a secure server) to verify the original scores provided by testers. If there were discrepancies between first and second scores, a core research team member who is an experienced clinician made the final scoring decision.

Scorers used an online spreadsheet to document the scoring process: the team would notate who second scored a test, their calculations of scores, any scoring discrepancies that were resolved, and any validity issues within a testing session. The scoring spreadsheet also contained formulas to automatically calculate raw scores to make the process more efficient. The second scorer documented the final scores in REDCap.

Scorers were encouraged to consult and communicate with the team whenever scoring questions or concerns arose.

Reliability. We computed Cronbach's alpha and split-half reliability for all of the standardized tasks administered in our study, except for one task where item-level information was not available from the publisher's website (DIBELS). Table 5.4 provides reliability coefficients for the current study and, for comparison, the coefficients provided from the publisher for each of the subtests. The reliability coefficients for the online administration of the subtests were comparable to those reported by the publishers and are considered to be within the acceptable-good range.

Measurement error. To further evaluate whether online administration of assessments introduced a measurement error, we calculated pairwise correlations among the standardized measures used in this study that overlapped with those administered for a different pre-pandemic in-person study in the lab (Lab Study A; Table 5.5). The comparison study (Lab Study A, Ozernov-Palchik et al., 2017) included 158 rising third-grade students with complete data for the relevant tests. Participants for this study were recruited from 21 schools in New England and represented a demographically similar sample to that of the current study (Figure 5.1). The correlation patterns among the variables in both studies were similar, suggesting that the same constructs were evaluated in the online version of the assessments as in the in-person version.

Previous in-person sample of 3rd graders *N*=158

	PPVT	CELF	KBIT	Blending Words	Memory for Digits
CELF	0.56***				
KBIT	0.30***	0.23**			
Blending Words	0.43***	0.39***	0.14		
Memory for Digits	0.46***	0.35***	0.27***	0.46***	
Nonword Repetition	0.56***	0.35***	0.15	0.57***	0.51***

Current Sample

	PPVT	CELF	KBIT	Blending Words	Memory for Digits
CELF	0.43***				
KBIT	0.49***	0.33***			
Blending Words	0.40***	0.28***	0.22**		
Memory for Digits	0.40***	0.23**	0.20**	0.29***	
Nonword Repetition	0.37***	0.31***	0.24**	0.40***	0.37***

Table 5.5: Pairwise correlations between six variables.

*Pairwise correlations between 6 variables for a previous in-person sample of third graders from our lab, and the current sample. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$*

Takeaways

The training process for administration and scoring of online assessments was more labor intensive than in-person studies, as there was an additional layer of developing tester competency with managing Zoom, engaging the child, and recording scores in an accessible way. Difficulties included connectivity issues and controlling for the

environment (i.e., background noise, distraction). The lack of control over the child's home environment posed some reliability and validity concerns, but the flexibility of online administration also allowed for a greater ability to adapt to children's and families' individual needs. Some children may have benefited from testing in their home environment, as testing in an unfamiliar location can lead to anxiety or stress.

For those planning to implement an online testing battery in an intervention study, we recommend setting up clear and detailed systems for documentation. The amount of digital documentation was greatly increased through adaptation for virtual administration. Materials and data should be organized in the most centralized and streamlined way possible to avoid confusion and misplacement of files. Not all stimuli and record forms can be easily adapted for online administration, and alternative methods (e.g., scanning the original form) may need to be considered depending on the assessment and availability of technology to testers and participants. It can be helpful to compile a document outlining each test, how it is administered, and links to any websites or documents needed for administration so the team has a centralized procedural document to follow.

We also recommend that before starting a new online study, researchers outline guidelines for addressing technical or environmental issues that inevitably arise (e.g., what to do if you are having trouble discerning the child's answer via Zoom). Technical and environmental factors cannot be eliminated when assessments are being administered virtually, but clear procedural guidance and detailed documentation during testing (e.g., noting the child's behavior and any technical issues) can help reduce reliability and validity concerns. Having an online, real-time messaging system (e.g., Slack) is also an essential tool to ensure the team is able to communicate

questions and concerns. Overall, the results for the standardized measures in the current study suggest equivalent effects of online testing to those of in-person testing, which are encouraging for the potential for future online intervention studies.

Intervention

Developmental researchers transitioning to a fully online intervention study will need to carefully consider how to adapt materials, train the research team (particularly if they are not located in the same place), and address difficulties that may be more likely to arise in online settings. In particular, intensive implementation of an intervention during the pandemic introduced new challenges related to privacy and disclosure. Finally, qualitative data on individuals' experiences participating in the study is important for identifying potential confounds and limitations, as well as for considering future scalability. We conclude this section by providing examples of feedback received from children and caregivers in our study.

Curriculum Adaptation

In our study, the Scaffolding Group received biweekly scaffolding sessions led by learning facilitators. For these sessions, we adapted an existing curriculum targeting oral language skills in elementary school children developed by the Language and Reading Research Consortium (LAARC; Goodwin, 2016). We added verbatim scripts for the learning facilitators to read to the children. Before each session, the learning facilitators adapted these scripts to the particular text they were working on with their child. The online format allowed learning facilitators to more easily follow a script than during face-to-face communication, thereby assuring greater fidelity of implementation. We also adapted materials that are designed for use by teachers in a

physical classroom to online administration. For example, we used the whiteboard feature in Zoom to draw and write words during the lesson. As part of their preparation for each lesson, the learning facilitators prepared slides with pictures of vocabulary words from the books. We embedded explicit instructions on how and when to utilize these virtual materials for each strategy. To avoid boredom and distraction, we incorporated activities to optimize child engagement during each lesson. All scaffolding sessions were recorded and stored on a secure server.

Learning Facilitator Training

We recruited and trained over 20 undergraduate students for the learning facilitator role during the study period. Students were interviewed and selected based on their experience and/or willingness to work with children and families, availability to meet consistently twice a week with their assigned participants via Zoom, and enthusiasm for the research. Our research team was ethnically and racially diverse, and many students were fluent in languages in addition to English. While over the summer most of the undergraduate students had full-time roles on the project, once the school year began, they had to juggle their work with their own courses and other responsibilities. Given the pandemic, many of the undergraduate students were not living on campus and completed their work from their own homes across the United States and in other countries. Our study team included many first-year students, students working in a lab for the first time, and students who did not come from a science background.

Learning facilitators underwent extensive training before being matched with participants to ensure implementation fidelity. First, we provided training in human subjects research, general strategies for working with young students, and background

literature on language/reading and summer interventions. Learning facilitators were also trained on the Learning Ally audiobook platform, and began reading the books used in our study. Because all of these training sessions were remote, learning facilitators could refer back to the recordings as needed. Next, we reviewed the scripts for each lesson with learning facilitators in group meetings. Learning facilitators paired up to practice each component of the lesson with each other (e.g., check in, vocabulary instruction, and scaffolding instruction). Each learning facilitator then recorded a full practice session which was reviewed by a member of the core research team. Learning facilitators received feedback on their recorded session, and those that required additional practice were asked to record new verification videos that implemented this feedback before being assigned participants. Undergraduate students who joined our team after the first summer were matched up with an experienced learning facilitator who served as a mentor and practice partner during training and beyond.

Crucially, training did not cease when learning facilitators began working with participants. All learning facilitators attended weekly meetings where they discussed their participants' progress and troubleshooted any issues. These issues ranged from how to properly implement specific strategies in the scaffolding curriculum, to how to communicate effectively with caregivers about scheduling, to how to respond to a child that shares difficult personal circumstances (see Child Disclosure below). Learning facilitators were encouraged to reach out to members of the research team any time they wanted to review a session and discuss strategies for working with a specific child, which was facilitated by the online nature of the study. A member of the research team also spot-checked session videos and provided feedback to learning facilitators as needed to ensure intervention fidelity. Finally, we cultivated an active community in a Slack channel, which allowed learning facilitators to post and answer questions

promptly. This multi-tiered network of support enabled our team of undergraduates to thrive in the remote research setting. Notably, in addition to all of their responsibilities as learning facilitators, undergraduates also filled numerous other roles on the project such as developing proximal assessment materials, transcribing language samples, communicating with caregivers, and assisting with data maintenance.

Online Intervention

Technical Challenges During Scaffolding Sessions. The biweekly scaffolding sessions over Zoom introduced challenges unique to the virtual setting. First, researchers were dependent on the capabilities of their own and the participant's internet connection and thus had to flexibly adapt when the connection was impaired. Many participants occasionally could not see or hear their learning facilitator during crucial parts of the session, or the learning facilitator could not discern what the participant was saying from the lagging audio. Learning facilitators took many steps to troubleshoot these issues while staying on Zoom. Turning cameras off, relocating closer to the Wi-Fi router, asking for a school-provided hotspot, and even using FaceTime or phone calls in tandem with Zoom helped mediate these issues. In a few cases, learning facilitators sent the session's materials to families ahead of time to print out or download so the child would not have to wait for webpages or screen-sharing to load. Learning facilitators also supported participants who had difficulty logging into or using the Learning Ally audiobook app by asking participants to share their screens and walking them through the setup.

The online setting also enabled children to multitask during sessions. For instance, there were numerous instances of participants attending sessions while siblings played

video games in the same room, while friends were over, or while simultaneously doing something else on the computer. To address these distractions, learning facilitators would ask, "Are you distracted right now? How can we fix that?" and having the child come up with potential solutions. These solutions included putting on headphones, moving to another room, or asking the people around them to quiet down.

Child disclosure. Disclosure of sensitive information occasionally came up during the testing and scaffolding sessions. In some cases this was prompted, as our study included parent and child questionnaires about experiences during the COVID-19 pandemic, negative feelings, and anxiety/depression. For example, a child disclosed that they thought about death "all the time" in response to a questionnaire item. We also anticipated that some scores on child self-report and parent-report anxiety/depression measures might fall in the clinically elevated range. In other cases, unprompted sensitive information was shared with researchers. For example, one child, when asked to use the vocabulary word 'evasive' in a sentence, said that they "used evasive action to avoid their mother hitting them." To address these expected and unexpected issues, we developed a detailed protocol for the research team to follow, overseen by a clinical psychologist who is a member of the research team. The psychologist checked the questionnaire data for red-flag indicators (supplemental protocol: https://osf.io/6urmx/) weekly. If there were indicators that met our criteria for concern (e.g., anxiety or depression scores that were in the clinically elevated range), she reviewed the pertinent data available and contacted the parents/guardians to alert them about the areas of concern and potentially suggest that they consider seeking a professional consultation for further guidance, if they had not already done so. In most cases, the parents/guardians were aware that their child was struggling emotionally (and many had already sought professional help or were in the process of doing so).

If a child indicated negative thoughts or feelings directly to a research team member during a session, the research team members were instructed to notify the psychologist immediately following the session. The psychologist would then follow up with the parents/guardians as necessary. We handled the incidence when a child came up with an example sentence about trying to avoid being hit by their parent differently. Although the role of researchers in mandatory reporting is debated, many states mandate researchers working with children to report suspicion of child abuse (Allen, 2009). Consequently, we called State Child and Family Services, where the family lives, and did an anonymous screening. Based on the information we provided, we were told that "it doesn't rise to the level of report." We continued to monitor the child, but nothing alarming came up during the subsequent sessions.

We learned from this study that particularly when frequently working with children directly in their homes or when collecting sensitive information, issues related to children's safety and wellbeing are likely to come up. We were fortunate to have a trained psychologist on our team who helped us develop a detailed protocol for dealing with these issues and who was responsible for communicating this information to families in a non-alarming but informative manner. Although not always mandated by the IRB, every study that involves children should include detailed procedures for handling sensitive information. Additionally, particularly for online studies that span several states, it is important to know which agency handles suspicions of potential abuse or neglect and what responsibilities researchers working with children have in that state.

Finally, it is important to support team members who may hear from children about difficult challenges they are facing. Most research assistants do not have mental health training, and thus may experience stress or other reactions to instances of child disclosure. Our learning facilitators were undergraduate students who themselves had been dealing with unprecedented challenges related to the pandemic. We addressed these potential challenges explicitly during training and through encouraging continuous communication within the team throughout the study, and by clearly indicating who to contact if such an issue arose. On our Slack channel and during weekly meetings, team members shared their experiences, debriefed, and coached each other on how to best respond to participants. In specific instances (described below), the clinical psychologist on our team provided one-on-one support to team members.

Qualitative Caregiver and Child Experiences

Child reflections. At the end of the eight weeks of meeting with learning facilitators, many participants did not want the study to end. When one learning facilitator started the last session with her student by saying, "Are you ready for our last lesson today?" the participant responded, "Yes, but I don't want it to be our last lesson," and ended up signing off the call by saying, "Okay, love you, see you, bye!" Another participant who always brought his favorite stuffed animal, Teddy, to the sessions remarked that, "Teddy is sad," when saying their goodbyes at their final meeting.

Many children reported enjoying the study experience, even if they did not enjoy their regular school-related activities or reading. During her final session, one student remarked "I hate school! School is evil." The learning facilitator said "Well, this is like

school and this was really fun!" to which the participant said, "This wasn't evil." One participant who had previously stated he did not enjoy reading told his parent at the 7-week mark: "You know what's so great about the audiobooks mum? It's that they're able to go into such more details than movies!" The parent expanded on this: "I cannot express the joy it brings me to hear my son starting conversations with me about stories he's read. Last week he wanted to recount some various storylines to me from books. To [say] that we've been enjoying the experience is an understatement. Thank you."

Many children also faced pandemic-related challenges that affected them during the course of the study. In addition to being out of school and having their social lives change, a few had family members who were directly affected by the virus. For instance, one participant was living with an uncle who had COVID-19. During one session, she told her learning facilitator, "People are in my house and it's difficult for me and my mom because, you know, my uncle is going to die. They want to help him, but they can't." One week later, during the routine check-in, the learning facilitator asked how she was doing and the participant said she was sad; "Yesterday, my uncle died. We saw him and, like, it's sad for me since I [have known] him since I was a kid. Me and my mom [were] crying." Her learning facilitator expressed her condolences, letting the child know that this is an extremely difficult time. She made sure to offer the participant an opportunity for breaks, instating a codeword of "rainbow sunshine." The learning facilitators adapted to meet the participants where they were at emotionally and mentally each session, knowing that the pandemic affected everyone's lives differently, and were generally a welcoming, consistent presence in the participants' lives for the duration of the study. Importantly, children participating in our research always come into our sessions with a variety of experiences. While the pandemic led to

more consistent challenges among our participants, these difficult experiences – death, illness, stress, financial insecurity – should always be on the research team's radar. At the end of the study, participants in the Scaffolding Group reported generally positive experiences (Table 5.6).

Caregiver reflections. At the end of the study, caregivers filled out a reflection survey about their experience in the study. In general, caregivers of children in the Scaffolding Group did not find it difficult for their child to have biweekly online meetings with their learning facilitator (Table 5.7).

Caregivers in both the Scaffolding Group and Audiobooks-only group likewise provided open-ended responses about their experiences in the study. Selected representative responses are included below (Table 5.8). Participants in the Audiobooks-only condition did not meet regularly with a learning facilitator, but they did receive weekly messages with updates on reading milestones and suggested book titles to read.

As reflected in these responses, caregivers in both groups had many positive experiences in the study. The remote learning environment fostered feelings of social isolation and loneliness for many children (as reflected in our surveys). In the Scaffolding Group, caregivers generally commented on interactions with the learning facilitators, and suggested that the connections forged between children and learning facilitators in our study may have helped ameliorate some of the negative socio-emotional consequences of the pandemic. This positive feedback is useful as we consider implementing future online interventions. In the Audiobooks-only Group, positive feedback focused on the reading experience and book selection.

Challenges were modest for both groups, and some challenges were not unique to the remote nature of the study. For instance, caregivers of children in the Scaffolding Group reported some difficulty finding time for sessions and getting their child to read the books, and some caregivers commented on the challenging nature of the vocabulary. In the Audiobooks-only Group, some caregivers noted that their child was not always interested in the recommended books. This group received the same book recommendations as the Scaffolding Group, but they did not discuss the books with a learning facilitator, which we hypothesized would impact their engagement. The Audiobooks-only Group also received only weekly updates; thus, they were unable to change books that did not interest them as easily as participants in the Scaffolding Group. Some caregivers also reported technical difficulties during and after the study. We relayed all technical issues to the audiobook company, and they worked with us and the caregivers to find solutions.

Takeaways

To properly measure intervention effects, we needed to ensure that both participants and learning facilitators were properly supported for an online intervention. Particularly for our learning facilitators, who had no previous experience implementing interventions, extensive training and open communication with supervisors and peers was critical. We found that weekly meetings and an internal study Slack channel provided opportunities for learning facilitators to learn from one another and troubleshoot issues. Consistent communication and chances to check-in were crucial since we could not share a physical lab space. Video recording of all sessions allowed

for ensuring fidelity of implementation and consistency across different learning facilitators and sessions.

The Scaffolding Group provided useful lessons for other researchers conducting studies with frequent online meetings. Researchers should expect some sessions to have distractions and technical difficulties; thus, it is important to have plans in place to ensure the fidelity of the study. Families reported only modest difficulties with study demands, and feedback from caregivers and children were overall positive. Indeed, many children felt comfortable sharing even highly personal information with their learning facilitators. Researchers should establish clear protocols for how to deal with sensitive information shared by children and families, particularly for studies that involve lots of online interactions.

Table 5.6: Child Experiences in Scaffolding Group.

How much did you like meeting with your learning facilitator?	
Not at all	1 (1.8%)
A little bit	3 (5.3%)
Sometimes	11 (19.3%)
A lot	42 (73.7%)

How often did you feel like you learned new words with your learning facilitator?	
Not at all	1 (1.8%)
A little bit	5 (8.8%)
Sometimes	11 (19.3%)
A lot	40 (70.2%)

Table 5.7: Caregiver Experiences in Scaffolding Group.

Was it challenging to get your child to meet with their learning facilitator?	
Not at all	50 (80.6%)
A little bit	9 (14.5%)
Sometimes	3 (4.8%)
A lot	0

Table 5.8: Caregiver Experiences in Scaffolding and Audiobooks-only Groups.

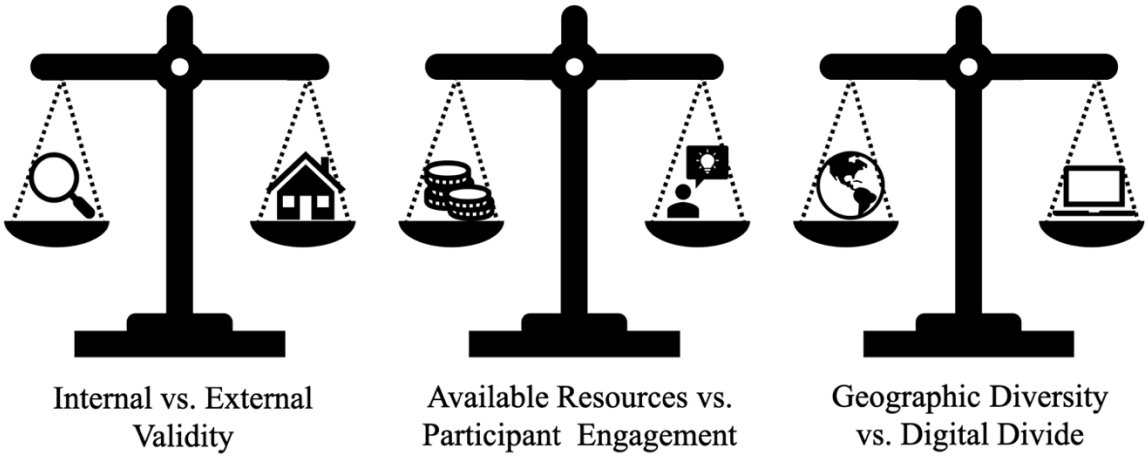
	Scaffolding Group	Audiobooks-only Group
What did your child enjoy most in this study?	"My child enjoyed all aspects of the study. He is proud to tell others that he is participating in a study. He is very excited to be paid by gift certificates. He loves how he can access any book of his choosing. He enjoyed the experience of meeting weekly and discussing the books with someone." "My son really enjoyed meeting with the learning facilitator and was sad to learn he would not be meeting with the facilitator anymore. He loved the books and the platform though I was hoping he would read more without me reminding him." "He enjoyed being introduced to books he may not have otherwise picked out to read. He also liked meeting with his facilitator. He is a social kid and the pandemic has been hard, so seeing [his Learning Facilitator] was a highlight of the week."	"He definitely enjoyed listening to the books that were recommended the best!!" "It allowed her to be independent with her nightly reading." "She enjoyed engaging with the tester. She enjoyed being able to pick her own book and listen on her own. This contributed to family conversations regarding the stories she listened too." "He really enjoyed the interviews and listening to/reading along w/ Learning Ally. I would like to continue it. He would often have siblings gathered around, reading too. ;)" "Es una experiencia bonita para los niños ,por que es una manera de leer sin leer osea escuchando ,es diferente pero me gusta,hasta la niña de segundo grado quería escuchar los libros ,me gusto mucho.gracias sigan asi ayudando a niños a que le den importancia a la lectura." Translation: "It is a beautiful experience for the kids because this way they can read with listening, it's different but I like it. Even my second grade daughter wanted to listen to the books. I enjoyed it a lot. Keep up the good work"
What did your child find most challenging in this study?	"She found the questions and vocabulary hard." "He is not used to listening to books and using the app required more setup time since he had to use his laptop, so it was something we had to remind him to do." "Finding time to read the books, especially without distraction" "Twice weekly meetings with the facilitator was a lot for our schedule"	"She did not like listening to books she had no interest in." "Trying to read/listen to the books she was not immediately interested in. I challenged her to try at least half of the book to see if it improved and she did not like that." "The second book didn't hold her interest" "Mostly technical problems" "Por las circunstancias pasa mucho tiempo conectado a algún dispositivo electrónico y aveces solo quería hacer otra cosa ,en circunstancias normales creo seria su actividad favorita."

"She sometimes did not want to stop what she was doing to attend scaffolding. Also wanted to socialize and share other things with Facilitator not fully focused on session"	Translation: "Because of the circumstances he spent a lot of time connected to an electronic device and sometimes he wanted to do something else. Under normal circumstances this might have been his favorite activity"

Discussion

We implemented a fully remote RCT intervention (final *N* = 255 third and fourth graders, ages 8-10 years) targeting children's language comprehension skills, which we described as a case study to explore various factors involved in conducting an online intervention study. We have summarized the challenges we faced, solutions we devised, and considerations for future research. Although our project represents a specific case study, and the implications should be considered carefully, we believe that the unique context of our study, its intensity and scale, and our diverse recruitment efforts allow us to derive 'lessons learned' that could be useful for others embarking on a similar project. We conclude by discussing what we believe to be the three main tradeoffs to think about when deciding whether and how to implement an online intervention study with a developmental sample (Figure 5.5).

Figure 5.5: Tradeoffs for online intervention studies with developmental populations.



Internal vs. external validity. An important goal of RCTs is to design and evaluate carefully controlled interventions that allow researchers to understand the precise causal mechanisms by which an intervention leads to learning gains. However, this can come at a cost – sometimes, the more controlled the intervention, the less likely it is to work in the "real world." As with any other type of study, an online RCT intervention requires researchers to consider tradeoffs between internal validity (how well the experiment tests what it is meant to test and is not influenced by other factors) and external validity (how well the experiment replicates in a natural environment).

Most developmental studies optimize internal validity by conducting studies in labs. These studies are well-poised to isolate the precise mechanism or phenomenon researchers are interested in studying. However, there are also drawbacks to in-lab studies that are particularly relevant for researchers interested in conducting RCTs. In-lab developmental studies typically rely on convenience samples, which tend to be homogenous, thereby limiting generalization to other populations (Bornstein et al., 2013). Furthermore, due to multiple practical considerations (e.g., space limitations,

transportation, scheduling issues), in-person studies tend to have smaller sample sizes than what is possible in online data collection. Finally, the ecological validity of such studies has been criticized - and the implications for what developmental processes look like in messy and unpredictable real-world settings, such as learning in a child's home, are limited (Lortie-Forgues & Inglis, 2019). Thus, while implementing an RCT study online in children's homes requires giving up some of the control of in-lab experiments and introduces additional noise, the tradeoff is that these studies can be more naturalistic and lead to increased sample diversity.

Especially important to consider for intervention studies is generalizability of effectiveness. On the other side of the spectrum from carefully controlled in-lab studies are large-scale educational RCT studies that implement interventions in schools and childcare settings. These studies tend to have higher external validity, but a side effect is increased noise. These studies often build on pilot studies that establish the value of a particular intervention under tightly-controlled conditions, but they tend to have small efficacy in these real-world settings (Lortie-Forgues & Inglis, 2019). There are many reasons for this. For example, school settings may be prohibitive of careful sample selection using stringent exclusion criteria (i.e., one child in a classroom receives the intervention while another child does not). Although there are design and statistical methods to overcome these issues (e.g., Regression Discontinuity Design; Lee & Munk, 2008), online intervention studies can bypass them altogether by working with eligible children in their own homes, which expands the pool of participants who are eligible to participate while also allowing the use of specific eligibility criteria and random group assignments. Similarly, it is more difficult to monitor and ensure implementation fidelity of programs when working in complex formal institutional environments such as schools, as compared to negotiating logistics with a child-

researcher duo. In our study, we were able to overcome these obstacles because we could closely monitor research activities via direct and continuous communication and video recording, and to document possible threats to validity during the various aspects of the study (e.g., background noise, child distraction, connectivity issues, implementation fidelity, etc.).

Thus, we suggest that the online implementation of intervention studies could improve the internal validity of such studies while maintaining their external validity. In online studies, the research team can operate within a well-controlled lab environment, while working with participants in natural, ecologically-valid settings. We discussed several potential threats to the validity of our study, such as background noise and technological challenges that could impact reliable data collection. Based on the comparison of the reliability scores for the current study and in-lab studies, however, online data collection resulted in equally reliable data collection, supporting the feasibility of maintaining internal validity in remote developmental research. The increased racial and socioeconomic diversity of the current sample, as compared to in-lab samples, suggests that we were able to achieve greater ecological validity. Furthermore, our study was conducted entirely in children's natural context – in their own homes – supporting its potential efficacy in real-world settings.

Available research resources vs. participant engagement. Implementing an RCT can be resource intensive – e.g., researchers' time, project budget, number of personnel – and often requires making decisions regarding how many resources to devote in order to maximize participant engagement and retention. Participant engagement can be measured across different levels (Matthews et al., 2011). Recruitment is one such measure that considers the reach of the study to the target population. Many

educational intervention studies rely on school partnerships for recruitment, which can be an effective strategy for recruiting a large number of children from diverse educational environments. However, establishing school partnerships requires substantial time and energy. The research team first has to clearly communicate the goals of the intervention and the benefits to that school's community in order to get buy-in from school leaders and educators. This process typically relies on existing relationships with schools and institutional familiarity, which might be more difficult for a new investigator to establish. Even when schools are interested in a potential partnership, the bureaucratic processes can be extensive before the study can get started. It can also be difficult to randomly assign students to conditions within a school because once a school is enthusiastic about an intervention, the school often wants all their students to be placed in the intervention condition.

On the other hand, many developmental science studies recruit participants directly through advertisements and social media (Hurwitz et al., 2017). Social media recruitment efforts can reach a wide pool of potential participants at a reasonably low cost. Our social media reach was extensive, reaching people from hundreds of different zip codes across the United States, but this required intentional targeted advertising. Based on our recruitment data, through school partnerships and social media, we successfully reached the participant demographic we set to recruit.

Enrollment, retention, and intervention adherence are additional types of engagement, each with its own set of challenges. Our enrollment and retention outcomes were less successful than our recruitment reach. Our final sample, although still very diverse, was not representative of the diversity in schools and communities we targeted in our recruitment. For example, household income eligibility for free/reduced lunch is

around $52,000. Although we targeted schools and communities with a high proportion of free/reduced lunch eligibility, we ended up with a median income with the $80,000-120,000 range. Thus, even though we allocated almost all of our recruiting budget and efforts to recruit lower-SES participants, our final enrollment was not skewed toward this demographic. Retention and intervention adherence represent two of the most critical factors to ensure the validity of intervention studies (Slack & Draugalis Jr, 2001) and are most difficult to achieve when working with disadvantaged communities. Ensuring participant engagement in such communities is resource-intensive, requiring a substantial recruitment budget, a large and well-trained research team, and attractive incentives for participation.

There is a large body of evidence from parenting programs targeting underserved communities that show how program-level factors (e.g., team member composition, level of family support provided) interact with participant factors (e.g., SES, job demands, perception of research, language barriers) in ensuring enrollment and retention (Hackworth et al., 2018; Whittaker & Cowley, 2012). Families, especially those from lower-SES backgrounds, are more likely to enroll and stay in a program, for example, if they have an experienced research liaison who supports them in identifying and overcoming barriers to participation (Hackworth et al., 2018; Rivas-Drake et al., 2016). Our full-time, bilingual coordinators were available to check in and assist families using preferred communication methods, and researchers assisted families with troubleshooting the apps for the intervention. Clear communication on research objectives and the theoretical foundation of the intervention is important for reducing perceptual barriers to participation (Barlow et al., 2003; Moran et al., 2004). Professionalism and experience of team members (Hackworth et al., 2018), as well as their representativeness of the target community (Gray, 2002), were additional factors

that ensured engagement. During our consent process, as well throughout the study, researchers were available to answer questions. We also hosted several information sessions for teachers and administrators in our partner district, as well as a bilingual (Spanish/English) session for parents at one of our partner schools. Intervention effects have been more significant in well-resourced studies, as compared to studies with fewer resources (Kim & Quinn, 2013). In general, across studies, there is an agreement that intervention programs targeting lower-SES communities require careful considerations of various factors that could affect direction of resources towards alleviating these barriers.

Online research may seem like a low-resource opportunity for obtaining larger, more diverse samples. With the advent of online platforms for developmental studies (e.g., Discoveries Online; Lookit), unmoderated research studies have become increasingly popular. Such studies, which allow participants to complete tasks on their own time and without the researcher's direct involvement, front-load their resources for design but require minimal resources for implementation. We caution, however, that families from underrepresented backgrounds may still face greater barriers to engaging in such studies than participants that are typically included in research studies, and we echo calls to actively work toward providing support and internet access for these populations (Lourenco & Tasimi, 2020; Sheskin et al., 2020). This is particularly pertinent for longitudinal and intervention studies that require substantial researcher moderation in order to be successful. Indeed, a similar online intervention during the pandemic that did not explicitly target a diverse sample based on SES ended up with almost all mothers with at least a 4-year college degree (Bambha & Casasola, 2021). We found that even children in school systems that did provide devices and internet access sometimes experienced technical difficulties in our study. Thus, while online

RCTs can remove certain resource constraints (such as space and travel compensation), researchers should expect to invest significant time and effort to achieve diverse samples and ensure their participation.

Geographic diversity vs. digital divide. Online study participation with children, although not always feasible, can significantly increase sample diversity by allowing easy access regardless of a family's geographic location and by minimizing caregivers' time commitment (Sheskin et al., 2020; Rhodes et al., 2020). This is particularly crucial for longitudinal studies that include multiple sessions and a significant time commitment. Online developmental studies have recruited more diverse samples than in-lab developmental studies (e.g., Scott et al., 2017; Scott & Schulz, 2017), including more geographically diverse samples (Bambha & Casasola, 2021). Our study recruited participants from 26 different states in the United States (Figure 5.2), and our sample was comparable to or better than our prior in-lab studies in terms of socioeconomic and racial diversity (Figure 5.1). However, the accessibility of online study participation is still challenging for many families (Lourenco & Tasimi, 2020). Prior to the start of the pandemic, almost a third of public K-12 students in the United States lacked adequate internet access and/or an adequate device for distance learning (Chandra et al., 2020). While some school systems provided children with computers and internet access to enable remote learning, many children still lack technology that would enable them to participate in an online intervention study. We unfortunately had to exclude interested families who lacked a computer or tablet at home due to our assessment battery. Furthermore, the "digital divide" – that is, the gap between people who have computer and internet access and those who do not – is not equally distributed across geographic boundaries and demographic groups (Van Dijk, 2020). 37% of students in rural communities in the United States lack adequate internet connectivity, compared

to 21% of students in urban environments (Chandra et al., 2020). Many of our participants struggled with internet connectivity issues and other technological challenges over the course of the study. Thus, it is important to take into account not only whether participants have access, but also whether they have complete access to these studies. In contrast, intervention studies that do not require the family to learn about the study and participate through their own technological platforms (such as most in-school interventions) allow researchers to ensure all participants in a constrained location can participate. Yet in-person interventions are not equally accessible to all geographic regions either - most of these studies take place near research institutions. One solution is to provide participants with the technology they need to participate in online research studies (Lourenco & Tasimi, 2020). Though adding additional costs to the study budget, providing devices with mobile data may lead to more representative samples as well as better data quality. For example, several large-scale projects have successfully deployed mobile devices loaded with educational content in rural locations in the US and around the world, like small villages in Ethiopia (Breazeal et al., 2016; Uchidiuno et al., 2018). This tradeoff may be worth the cost, particularly for home-based intervention studies. Online studies allow for geographic diversity of the research team as well. Our study team worked from multiple time zones, which allowed us to accommodate participants from across the United States. This also opens up the possibility for recruiting community members to be part of the research team. This type of participatory research may lead to higher recruitment, retention, and validity of intervention studies (Levac et al., 2019).

Conclusion

In response to the COVID-19 pandemic we conducted a scalable online RCT intervention study with children from diverse backgrounds across the United States. In this paper, we summarized the challenges we encountered and the tradeoffs to consider when implementing such studies. Despite possible threats to the internal validity of our study, difficulties in reaching demographically diverse populations, and resource-exhaustive efforts to support participant engagement and retention, we were able to conduct a study that provided educational support during a challenging time for both children and their caregivers. With the aforementioned considerations and tradeoffs in mind, we believe that fully remote intervention studies are a worthwhile endeavor for developmental researchers, and we expect to see more of them in the future.

Acknowledgements

We are grateful for the partnership with Learning Ally, for helpful suggestions from James Kim and Tiffany Hogan, and for funding support from the Chan Zuckerberg Initiative for the Reach Every Reader project (https://www.gse.harvard.edu/reach-every-reader). We would also like to thank our undergraduate research assistants: Tolu Asade, Cherry Wang, Sophia Angus, Bhuvna Murthy, Maycee McClure, Sehyr Khan, Hilary Zen, Sarah Abodalo, Elizabeth Carbonell, Shruti Das, Erika Leasher, Yoon Lim, Emmi Mills, Zoë Elizee, Camille Uldry, Shelby Laitipaya, Alexis Cho, Gabriella Aponte, Harley Yoder, Dana Osei, Niki Kim, Joy Bhattacharya; our testers: Amanda Miller, David Bates, Ross Weissman, Joohee Baik, June Okada, William Oliver, Harriet Richards; and our additional team members: Isaac Treves, Cindy Li, Kristen Wehara, Brooke Goldstein, Ada Huang. We are also grateful to the families for their time and participation.

References

Allen, B. (2009). Are researchers ethically obligated to report suspected child maltreatment? A critical analysis of opposing perspectives. *Ethics & Behavior*, *19*(1), 15–24.

Bambha, V. P., & Casasola, M. (2021). From Lab to Zoom: Adapting training study methodologies to remote conditions. *Frontiers in Psychology*, *12*.

Barlow, J., Coren, E., & Stewart-Brown, S. (2003). Parent-training programmes for improving maternal psychosocial health. *Cochrane Database of Systematic Reviews*, *4*.

Barton, E. E., & Fettig, A. (2013). Parent-implemented interventions for young children with disabilities: A review of fidelity features. *Journal of Early Intervention*, *35*(2), 194–219.

Bornstein, M. H., Jager, J., & Putnick, D. L. (2013). Sampling in developmental science: Situations, shortcomings, solutions, and standards. *Developmental Review*, *33*(4), 357–370.

Breazeal, C., Morris, R., Gottwald, S., Galyean, T., & Wolf, M. (2016). *Mobile devices for early literacy intervention and research with global reach*. 11–20.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*(6), 2156–2160.

Chandra, S., Chang, A., Day, L., Fazlullah, A., Liu, J., McBride, L., Mudalige, T., & Weiss, D. (2020). Closing the K–12 digital divide in the age of distance learning. *Common Sense and Boston Consulting Group: Boston, MA, USA*.

Dunn, L. M., & Dunn, D. M. (2007). Peabody picture vocabulary test–fourth edition (PPVT-4). *Circle Pines, MN: AGS*.

Fixen, D., Naoom, S., Blase, K., Friedman, R., & Wallace, F. (2005). Implementation Research: A Synthesis of the Literature. Tampa, Fl: University of South Florida: The National Implementation Research Network.

George, F., & Mallery, M. (2003). Quantitative method for estimating the reliability of data. *Western Machinegun University*.

Gillen, N. A., Siow, S., Lepadatu, I., Sucevic, J., Plunkett, K., & Duta, M. (2021). Tapping into the potential of remote developmental research: Introducing the OxfordBabylab app.

Good, R. H., Kaminski, R. A., Smith, S., & Laimon, D. (2002). *Dynamic indicators of basic early literacy skills: DIBELS*. Dynamic Measurement Group.

Goodwin, A. P. (2016). Effectiveness of word solving: Integrating morphological problem-solving within comprehension instruction for middle school students. *Reading and Writing*, *29*(1), 91–116.

Gray, B. (2002). Emotional labour and befriending in family support and child protection in Tower Hamlets. *Child & Family Social Work*, *7*(1), 13–22.

Hackworth, N. J., Matthews, J., Westrupp, E. M., Nguyen, C., Phan, T., Scicluna, A., Cann, W., Bethelsen, D., Bennetts, S. K., & Nicholson, J. M. (2018). What influences parental engagement in early intervention? Parent, program and community predictors of enrolment, retention and involvement. *Prevention Science*, *19*(7), 880–893.

Harris, P. A., Taylor, R., Minor, B. L., Elliott, V., Fernandez, M., O'Neal, L., McLeod, L., Delacqua, G., Delacqua, F., & Kirby, J. (2019). The REDCap consortium: Building an international community of software platform partners. *Journal of Biomedical Informatics*, *95*, 103208.

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377–381.

Heinrichs, N., Bertram, H., Kuschel, A., & Hahlweg, K. (2005). Parent recruitment and retention in a universal prevention program for child behavior and emotional problems: Barriers to research and program participation. *Prevention Science*, *6*(4), 275–286.

Hurwitz, L. B., Schmitt, K. L., & Olsen, M. K. (2017). Facilitating development research: Suggestions for recruiting and re-recruiting children and families. *Frontiers in Psychology*, *8*, 1525.

Kaufman, A. S. (2004). Kaufman brief intelligence test–second edition (KBIT-2). *Circle Pines, MN: American Guidance Service*.

Kim, J. S., & Quinn, D. M. (2013). The effects of summer reading on low-income children's literacy achievement from kindergarten to grade 8: A meta-analysis of classroom and home interventions. *Review of Educational Research*, *83*(3), 386–431.

Kizilcec, R. F., Reich, J., Yeomans, M., Dann, C., Brunskill, E., Lopez, G., Turkay, S., Williams, J. J., & Tingley, D. (2020). Scaling up behavioral science interventions in online education. *Proceedings of the National Academy of Sciences*, *117*(26), 14900–14905.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253.

Lee, H., & Munk, T. (2008). Using regression discontinuity design for program evaluation. 3–7.

Levac, L., Ronis, S., Cowper-Smith, Y., & Vaccarino, O. (2019). A scoping review: The utility of participatory research approaches in psychology. *Journal of Community Psychology*, *47*(8), 1865–1892.

Lingwood, J., Levy, R., Billington, J., & Rowland, C. (2020). Barriers and solutions to participation in family-based education interventions. *International Journal of Social Research Methodology*, *23*(2), 185–198.

Lorenc, T., Petticrew, M., Welch, V., & Tugwell, P. (2013). What types of interventions generate inequalities? Evidence from systematic reviews. *J Epidemiol Community Health*, *67*(2), 190–193.

Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher*, *48*(3), 158–166.

Lourenco, S. F., & Tasimi, A. (2020). No participant left behind: Conducting science during COVID-19. *Trends in Cognitive Sciences*, *24*(8), 583–584.

Manz, P. H., Hughes, C., Barnabas, E., Bracaliello, C., & Ginsburg-Block, M. (2010). A descriptive review and meta-analysis of family-based emergent literacy interventions: To what extent is the research applicable to low-income, ethnic-minority or linguistically-diverse young children? *Early Childhood Research Quarterly*, *25*(4), 409–431.

Matthews, E. C., Fox, S., Hackworth, N., Kitanovski, M., & Vista, A. (2011). The Parenting Research Centre. We wish to acknowledge the valuable contributions and support of: Iris Crook, Community Development Worker, Family & Children's Services, Yarra Ranges Shire; Georgina Devereaux, Playgroup Support and Development Officer, Frankston City Council—.

Moran, P., Ghate, D., Van Der Merwe, A., & Policy Research Bureau. (2004). *What works in parenting support?: A review of the international evidence*. DfES Publications London.

Nicholson, L. M., Schwirian, P. M., Klein, E. G., Skybo, T., Murray-Johnson, L., Eneli, I., Boettner, B., French, G. M., & Groner, J. A. (2011). Recruitment and retention strategies in longitudinal clinical studies with low-income populations. *Contemporary Clinical Trials*, *32*(3), 353–362.

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of Experimental Child Psychology*, *162*, 31–38.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Review of Educational Research*, *78*(1), 33–84.

*Olson, H. A., *Ozernov-Palchik, O., Arechiga, X. M., Wang, K. L., & Gabrieli, J. D. E. (in prep). Effects of remote voluntary audiobook randomized controlled trial intervention on children's language skills. *Manuscript in Preparation.*

Ozernov-Palchik, O., Norton, E. S., Sideridis, G., Beach, S. D., Wolf, M., Gabrieli, J. D., & Gaab, N. (2017). Longitudinal stability of pre-reading skill profiles of kindergarten children: Implications for early screening and theories of reading. *Developmental Science*, *20*(5), e12471.

Pollack, C., Wilmot, D., Centanni, T., Halverson, K., Frosch, I., D'Mello, A., Romeo, R. R., Imhof, A., Capella, J., & Wade, K. (2021). *Anxiety, motivation, and competence in mathematics and reading for children with and without learning difficulties*.

Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., Benitez, J., & Ocampo, J. D. (2020). Advancing developmental science via unmoderated remote research with children. *Journal of Cognition and Development*, *21*(4), 477–493.

Rivas-Drake, D., Camacho, T. C., & Guillaume, C. (2016). Just good developmental science: Trust, identity, and responsibility in ethnic minority recruitment and retention. *Advances in Child Development and Behavior*, *50*, 161–188.

Scott, K., Chu, J., & Schulz, L. (2017). Lookit (Part 2): Assessing the viability of online developmental research, results from three case studies. *Open Mind*, *1*(1), 15–29.

Scott, K., & Schulz, L. (2017). Lookit (part 1): A new online platform for developmental research. *Open Mind*, *1*(1), 4–14.

Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., Fei-Fei, L., Keil, F. C., Gweon, H., & Tenenbaum, J. B. (2020). Online developmental science to foster innovation, access, and impact. *Trends in Cognitive Sciences*, *24*(9), 675–678.

Slack, M. K., & Draugalis Jr, J. R. (2001). Establishing the internal and external validity of experimental studies. *American Journal of Health-System Pharmacy*, *58*(22), 2173–2181.

Uchidiuno, J., Yarzebinski, E., Madaio, M., Maheshwari, N., Koedinger, K., & Ogan, A. (2018). *Designing appropriate learning technologies for school vs home settings in tanzanian rural villages*. 1–11.

Van Dijk, J. (2020). *The digital divide*. John Wiley & Sons.

Veinot, T. C., Mitchell, H., & Ancker, J. S. (2018). Good intentions are not enough: How informatics interventions can worsen inequality. *Journal of the American Medical Informatics Association*, *25*(8), 1080–1088.

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (1999). *Comprehensive test of phonological processing: CTOPP*. Pro-ed Austin, TX.

Wechsler, D. (2011). WASI-II: Wechsler abbreviated scale of intelligence. PsychCorp.

Whittaker, K. A., & Cowley, S. (2012). An effective programme is not enough: A review of factors associated with poor attendance and engagement with parenting support programmes. *Children & Society*, *26*(2), 138–149.

Wiig, E. H., Secord, W. A., & Semel, E. (2013). *Clinical evaluation of language fundamentals: CELF-5*. Pearson.

"Inconceivable!"

"You keep using that word. I do not think it means what you think it means."

— William Goldman, ***The Princess Bride***

Chapter 6 : Preliminary effects of listening to audiobooks with instructional support on children's vocabulary

**This chapter includes preliminary and exploratory analyses that differ from the preregistered analysis plan, but that relate to the overall themes of this thesis. These will not be the final set of analyses submitted for publication, and thus should be interpreted accordingly.*

Some of the following chapter appeared in:

†Olson, H., †Ozernov-Palchik, O., Arechiga, X., Wang, K., Dieffenbach, J., & Gabrieli, J. D. E. (2022). A Remote Randomized Controlled Trial Audiobook Intervention. *International Mind, Brain, & Education Society Conference 2022*. Poster: Montreal, Canada.
†*Authors share joint first authorship.*

Abstract

Reading books is an opportunity for children to encounter more complex vocabulary and language than they are exposed to in everyday speech. However, children vary widely in the amount of time they spend reading, and those who struggle to read are often less motivated to spend their free time reading. To determine whether removing the "decoding" part of reading as a potential barrier would improve children's vocabulary and other language skills, we designed a randomized controlled trial intervention study in which children listened to audiobooks along with text, either alone or with scaffolded instructional support. Third and fourth-grade students (N=311, ages 8.0-10.8 years) were randomly assigned to Audiobooks+Scaffolding, Audiobooks-Only, or Mindfulness (control group) for 8 weeks. Using book-specific "proximal measures" of receptive and expressive vocabulary, we found that poor readers only

showed pre-to-post gains in receptive and expressive vocabulary in the Audiobooks+Scaffolding group. These preliminary results suggest that listening to audiobooks paired with scaffolded instructional support may be a promising approach to support vocabulary learning in struggling readers.

Introduction

Children's linguistic experiences have a profound effect on their language development. This is obviously true in infancy, yet the impact of linguistic input does not stop once children gain fluency in their native language. In particular, people will continue to acquire new vocabulary through adulthood. A key source of input for vocabulary, as well as syntactic structures and complex language, is reading (Biber, 1991; Montag et al., 2015; Montag & MacDonald, 2015), and indeed, reading experience has been linked to the development of language skills (Acheson et al., 2008; Duff et al., 2015). However, there is immense variation in the time children spend reading depending on their reading skills. A study from 1988 estimated, based on a composite measure of reading speed and free time spent reading, that 5th grade students at the 90th percentile may read as many words in a little over a week as a child at the 10th percentile might read in a year (Anderson et al., 1988). Even though these exact numbers may be out of date due to technological advances and shifts in how children spend their free time, variations in the amount of time spent reading can strikingly impact children's language experiences.

Intrinsic motivation to read has been positively associated with reading achievement (e.g., Troyer et al., 2019). Children who struggle to read are generally less motivated to read (Melekoğlu & Wilkerson, 2012) and choose to read less (van Bergen et al., 2018), and many children struggle with reading comprehension: according to the 2022 National Assessment of Education Progress, only 33% of fourth graders in the United

States are reading at or above a proficient level (*NAEP Reading: Reading Highlights 2022*). Even if there were no effect of reading skills on time spent reading, children reading below grade level are less likely to be exposed to the type of linguistic content found in grade-level books.

If struggling to read begets less motivation to read, and less reading means that these children are accessing different linguistic input and background knowledge, then one prudent target of intervention may be to remove reading as a barrier to accessing more complex language input. In this study, we asked: can we change the linguistic environment, and thereby impact reading comprehension and component processes?

According to the Simple View of Reading (Gough & Tunmer, 1986; Hoover & Gough, 1990), successful reading comprehension requires both (1) accurate word reading, and (2) good language comprehension, which includes vocabulary knowledge and language skills that support the understanding of spoken language. Multiple studies have shown that word reading and language comprehension are dissociable, and that both contribute to successful reading (e.g., Catts et al., 2006; Kendeou et al., 2009; Scarborough, 2001; Tilstra et al., 2009).

In this study, we targeted language comprehension through both implicit and explicit intervention strategies in a randomized controlled trial (RCT) intervention. The implicit strategy was to expose children to more complex language by encouraging them to voluntarily listen to audiobooks. Audiobooks remove decoding as a barrier to reading, exposing children to narrative structure, new vocabulary, and syntactic complexity that are crucial to reading comprehension, but are not always found in everyday speech. Audiobooks and text-to-speech or read-aloud tools may support comprehension,

particularly in younger children and children with reading difficulties (Singh & Alexander, 2022; Wood et al., 2018), but overall, this approach is underexplored. Thus, we hypothesized that listening to audiobooks that introduce new vocabulary and appropriately challenging content would improve children's language skills.

The explicit strategy to target language comprehension was to provide children with scaffolding instructional support. In addition to listening to audiobooks, we randomly assigned one group of children to receive one-on-one instructional support during the intervention (Language and Reading Research Consortium, 2016; Language and Reading Research Consortium, Arthur, et al., 2016; Language and Reading Research Consortium et al., 2014). Studies on voluntary reading interventions have revealed some efficacy (Kim & Quinn, 2013), but the impact of voluntary reading on reading achievement has been questioned (National Reading Panel (U.S.), 2000). Some previous voluntary summer reading interventions for elementary school students have incorporated structured, scaffolded lessons targeting reading comprehension and related skills in addition to providing access to reading level-appropriate and interest-matched books. Adding these scaffolds resulted in greater reading achievement and reduced summer learning loss (Kim & White, 2008; White & Kim, 2008). Our scaffolding sessions also included explicit vocabulary instruction, as vocabulary instruction has been shown to positively impact reading comprehension, especially when there are multiple exposures to words (e.g., Elleman et al., 2009; Stahl & Fairbanks, 1986), though transfer effects may be limited (Wright & Cervetti, 2017).

Ultimately, the goal of this intervention was to improve reading comprehension. Yet reading comprehension is a complex, multifaceted process – thus, trying to measure reading comprehension is fraught (Paris & Stahl, 2005). Measures of reading

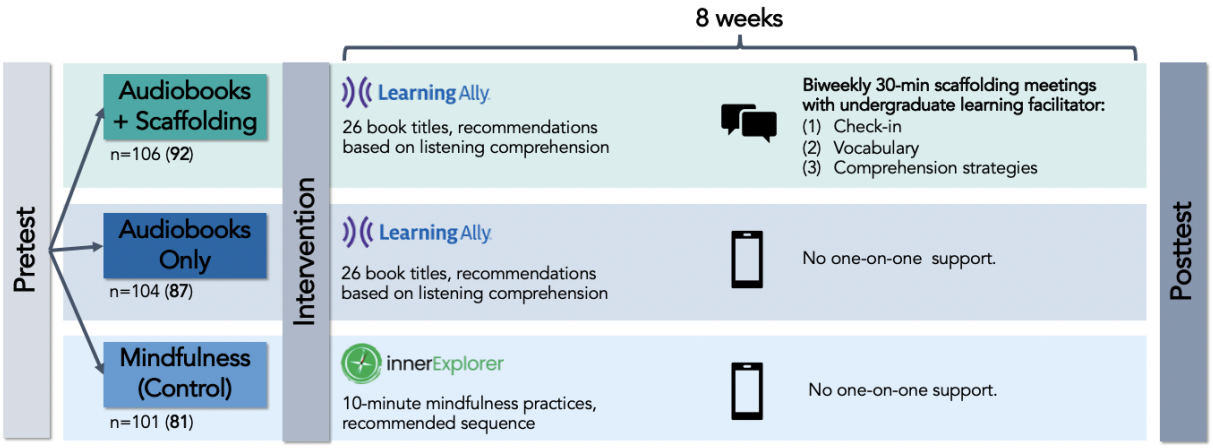
comprehension are influenced by how interesting or familiar the material is (Baldwin et al., 1985; Shnayer, 1968), as well as whether questions are answered within or at the end of the text (Guerreiro et al., 2022). A slightly more defined construct is vocabulary, which can be subdivided into receptive vocabulary knowledge (i.e., the words one can recognize) and expressive vocabulary knowledge (i.e., the words one can explain). Receptive and expressive vocabulary can be differentially impacted by book reading interventions in young children, depending on the type of support (e.g., Sénéchal, 1997), though the type of support for vocabulary learning varies by study, and oftentimes relatively few words are learned (Wasik et al., 2016). Because the scaffolding instruction includes explicit vocabulary instruction, we expected that this audiobook intervention might particularly affect vocabulary skills. In addition to standardized measures of vocabulary, we also created "proximal" vocabulary tests, for each book, that included words that appeared in the books children read, as we hypothesized that proximal measures may be more likely to detect effects from this relatively short intervention (e.g., Solari et al., 2020; Vadasy et al., 2015). We selected words based on the average age of acquisition, aiming for words that would be unfamiliar but useful in broader contexts (Beck et al., 2002)[25]. This chapter presents preliminary, exploratory results from both the standardized and proximal vocabulary measures as indicators of intervention efficacy.

[25] Beck, Mckeown, and Kucan describe three "tiers" of vocabulary: Tier One words are basic words that are typically learned in conversation without explicit instruction, Tier Three words are low frequency and domain-specific, and Tier Two words are more useful and versatile words that are typically found in written language and are rarely used in everyday conversation. We tried to choose Tier Two words for vocabulary instruction and assessment.

Methods

Study Overview: This randomized controlled trial (RCT) study was conducted completely remotely during the COVID-19 pandemic, with data collection beginning in mid-summer 2020 and concluding in spring 2022. The study consisted of three main phases: Pretest, Intervention, and Posttest (**Figure 6.1**). Children were compensated $20 per hour for all pretesting and posttesting sessions, and caregivers were additionally compensated $5 per survey for completing a total of 10 surveys at the beginning and end of the intervention period. Study procedures were approved by MIT's Committee on the Use of Humans as Experimental Subjects. Detailed experimental procedures and considerations for online RCTs with developmental populations were previously published in Ozernov-Palchik, Olson, et al., 2022 (**Chapter 5**).

Figure 6.1: Study design.



Schematic of study design. Shows number of participants randomized into each group, and the number who completed some posttesting in parentheses.

Participants: 311 children (ages 8.0-10.8 years, mean(SD)= 9.5(.6) years) participated in the study. Children were enrolled in the 3rd or 4th grade when they began the study. Participants were randomly assigned to one of the intervention conditions: Audiobooks+Scaffolding (N=106), Audiobooks-Only (N=104), or Mindfulness (N=101). Of these participants who were assigned to a group and began the intervention, 260 completed at least some posttesting (84% retention overall; N=92 for Audiobooks+Scaffolding, N=87 for Audiobooks-Only, and N=81 for Mindfulness completed at least some posttesting). Participants were primarily recruited via school partnerships and online advertising (see Ozernov-Palchik, Olson, et al., 2022 for additional details). To be eligible for the study, children had to be fluent in English, have a parent or guardian that spoke English or Spanish, have normal or corrected to normal hearing, have no diagnosis of autism spectrum disorder, and have a nonverbal reasoning standard score of 80 or above on the Kaufman Brief Intelligence Test administered during pretest (KBIT-2; Kaufman & Kaufman, 2004).

Intervention: Following the initial pretest, participants were assigned to an intervention group for an approximately 8-week period. The start dates were rolling, such that a participant's start date was always on a Monday.

Audiobooks-Only: Participants were given access to the *Learning Ally Audiobook Solution* platform, which contains a library of audiobooks along with text that can be accessed via computer, smartphone, or tablet. As the book is read aloud, the words are highlighted on the screen. Each participant's account had a set of recommended books selected based on their listening comprehension level (see Supplementary Table 1 for lists of books). The goal was to create book lists in children's zone of proximal development (Vygotsky & Cole, 1978) - that is, slightly challenging based on their

current listening comprehension level, but that they could access with support. There were three 'tracks' of curated books, and all tracks included fiction and nonfiction options, as well as diverse characters. Participants were asked to listen to their books for approximately 90 minutes per week. They were instructed to listen to one book at a time, and were typically given a choice between two books each time they finished a book, with the exception of the first and last book which we aimed to standardize between all children at a particular level.

Audiobooks+Scaffolding: The intervention was identical to the Audiobook-Only group, with the addition of biweekly one-on-one "scaffolding sessions" focused on vocabulary and reading comprehension. Each participant was assigned to an undergraduate student who served as their "Learning Facilitator" for the study. Participants in the Audiobooks+Scaffolding Group met with their Learning Facilitator one-on-one via Zoom for two 30-minute sessions per week for the duration of the intervention. During these sessions, Learning Facilitators were asked to: (1) check in about reading progress, identify barriers to reading, and brainstorm suggestions as needed, (2) teach two vocabulary words from the text, and (3) teach and review a reading comprehension strategy using a lesson plan (based on Language and Reading Research Consortium curriculum, (Goodwin, 2016; Language and Reading Research Consortium), see **Supplementary Table 2** for overview). Learning Facilitators logged notes for each session, and all sessions were recorded with permission from caregivers and children.

Mindfulness: Each participant completed mindfulness exercises provided by Inner Explorer. Participants were instructed to complete five 10-minute mindfulness practices per week for the intervention period.

Testing: 2-3 pretest sessions were conducted via zoom by experimenters blind to group assignment. Assessments were administered in a priority order, with listening comprehension, nonverbal reasoning, and vocabulary measures administered during Pretest 1, and other language and reading measures, as well as book 1 measures, administered during Pretest 2. Pretest 3 was completed after the first book was read (in the two audiobooks groups), or approximately 2 weeks into the intervention period (for the Mindfulness group), during which book 1 posttest measures were administered, as well as any remaining measures that were not completed during the initial pretest sessions. 1-3 posttest sessions were conducted via zoom by experimenters blind to group assignment (until the final questionnaire). Assessments were again administered in a priority order, with final book measures administered during the first posttest session.

Standardized Vocabulary Assessments: Assessments were administered via zoom during both the Pretest and Posttest sessions. These preliminary exploratory analyses will focus on only the vocabulary measures: our receptive vocabulary measure was the Peabody Picture Vocabulary Test (PPVT-5; Dunn, 2018), and our expressive vocabulary measure was the Vocabulary subtest of the Wechsler Abbreviated Scale of Intelligence (WASI-II; Wechsler, 2011).

Proximal Vocabulary Assessments: In addition to the standardized assessments, we developed book-specific "proximal measures" of vocabulary. Given the brief nature of the intervention, it was possible that gains made during the intervention period would not be picked up by the standardized assessments. For instance, listening to audiobooks may boost vocabulary learning as we hypothesized, but if the words children learned from the books did not show up on the standard assessments, we

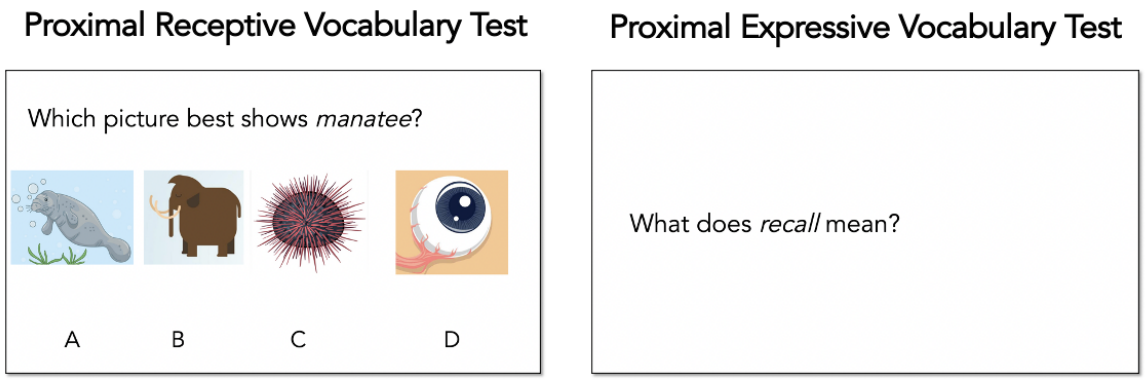
would not be able to measure these gains. Thus, we developed proximal vocabulary tests for each of the 26 titles.

For each book, we used frequency (how many times each word appears in the book), age of acquisition (Brysbaert & Biemiller, 2017; Kuperman et al., 2012), and concreteness ratings (Brysbaert et al., 2014) to select 16 target words per book, as well as 4 "easy" words from the book matched on frequency and 4 "non-book" words matched on age of acquisition that did not appear in the book. We ensured that items did not repeat on books within the same track. To do this, we selected all words in the book within an age of acquisition bin based on their assigned track (6-8 years for Track 1, 8-10 years for Track 2, and 10-12 years for Track 3). Next we sorted all the words from high to low on frequency within the book, and selected words that would work for receptive items (based on concreteness ratings and experimenter judgment) and expressive items (based on experimenter judgment).

Each proximal vocabulary test contained receptive items (4-choice multiple choice with pictures, analogous to the PPVT; **Figure 6.2 left**) and expressive items (requiring students to define a word, scoring procedures based on the WASI-2 vocabulary subtest; **Figure 6.2 right**). Words with higher concreteness ratings were selected for the receptive vocabulary items so that we could represent them pictorially. *Proximal Receptive Vocabulary Test*: For the receptive items, we included one semantic foil and one phonological foil, plus an additional word matched on age of acquisition. For these analyses, the Proximal Receptive Vocabulary score is the sum of the number of correct responses to the 10 target words (maximum score = 10). *Proximal Expressive Vocabulary Test*: For the expressive items, participants were asked to define the target word. Responses were scored from 0-2 using a rubric, such that 0=No Knowledge

(e.g., no response, inappropriate use in phrase/sentence, inappropriate definition, restatement, phonological manipulation), 1=Incomplete Knowledge (e.g., appropriate use in phrase/sentence (uses it without really defining the word), vague/imprecise definition, imprecise synonym, or 1-H: homophone is defined rather than the target word), and 2=Complete Knowledge (e.g., precise use in phrase/sentence, precise definition/synonym). Independent experimenters used a subset of participant responses to create rubrics for each word in each book, with examples of what would constitute a 0, 1, and 2 response, then independent experimenters scored a subset of new tests to measure reliability, then new experimenters scored the remaining responses. There were 6 target words per book (maximum score = 12).

Figure 6.2: Proximal vocabulary assessments.



Example items from the Proximal Vocabulary Test for the book Crenshaw. ***Left****: Receptive item. Options include the target word (A; manatee), a phonological foil (B; mammoth), a semantic foil (C; sea urchin), and an age of acquisition-matched word (D; iris).* ***Right****: Expressive item. Items were read aloud to the child and displayed on the screen.*

Book 1 pretest was administered at Pretest 2, and book 1 posttest was administered at pretest 3 (which was scheduled after children in the audiobooks group finished book 1,

or about 2 weeks after pretest 2 for the Mindfulness group). Final book pretest was administered at pretest 2 or 3, and the final book posttest was administered at posttest 1. The tests were the same for pretest and posttest for each book. In some cases, children in the audiobook groups did not read the intended final books, and were instead post-tested on the final book they completed. Tests were administered before and after each book for the Audiobooks+Scaffolding group, so if they read a different final book, then they were prettested during a scaffolding session prior to reading the final book. The Mindfulness group did not read the books, so we picked the tests based on what would be their assigned track based on their listening comprehension and receptive vocabulary scores. For the results reported here, scores are from the final book, as pretesting was completed at the beginning of the intervention period[26] and posttesting was completed at the end of the intervention period.

Additional Measures: In addition to vocabulary, children were also assessed on measures of listening comprehension, reading comprehension, reading fluency, and working memory. For additional standardized measures and their validity for online administration, see Chapter 5 (Ozernov-Palchik, Olson, et al., 2022). For exploratory analyses, we used the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Passage Reading Fluency subtest score to define Good Readers (>20th percentile) and Poor Readers (≤20th percentile, considered "at risk" for reading difficulty) (Good & Kaminski, 2002). Children were also assessed via a proximal comprehension test at the end of each book, using questions developed by educators at *Learning Ally*. These questions were administered to the Audiobooks+Scaffolding group after each book, and to the Audiobooks-Only group for the first book and final book. Finally, all children

[26] With the exception of some scaffolding participants who did not reach the final book.

also responded to an open-ended prompt ("Tell us about a book you read recently") at the beginning and end of the study to assess expressive language. Responses are being transcribed and analyzed for vocabulary, mean length of utterance, and other linguistic features. Finally, parents filled out surveys about their child's background and home learning environments, with a specific focus on the impact of COVID-19 on the learning environment. Children likewise responded to surveys about their learning experiences and reading motivation at the beginning and end of the study.

Preliminary Exploratory Analyses: Exploratory analyses on the preliminary data focused on the vocabulary measures. We asked whether there was an interaction between group assignment, reading fluency (binary factor based on DIBELS percentile score), and time (pre-to-post test). Specifically, we used a linear mixed effects model, with participant as a random effect: score~age+ReadingFluency*Group*Time+(1|ParticipantID). We then performed post-hoc Tukey tests to examine how pre-to-post change varied by group and reading fluency. Note that the preregistered analyses to test for intervention effects will use imputed data (assuming missing data at random) in a confirmatory factor analysis model, taking additional variables into account.

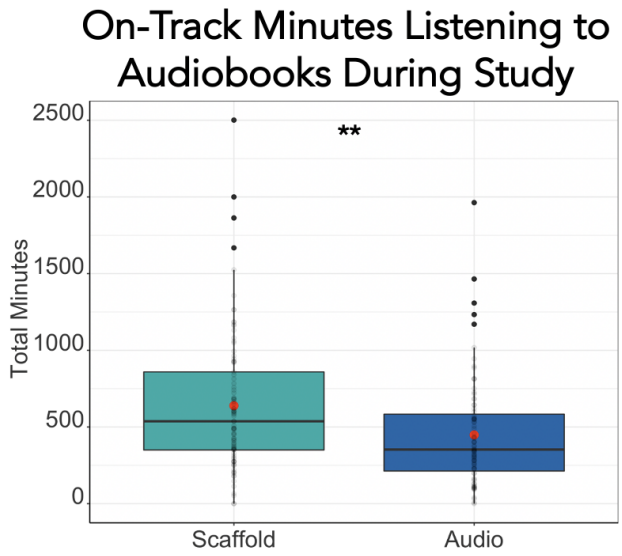
Preliminary Results

One-on-one scaffolding increased time spent listening to recommended audiobooks.

To explore whether the one-on-one support increased intervention compliance, we calculated the total minutes spent listening to the recommended audiobooks on the *Learning Ally* platform for the duration of the study. The Audiobooks+Scaffolding

group spent significantly more time listening to their books (M(SD)=638.9(448.6) minutes, range 0-2501 minutes, *NA*=15) than the Audiobooks-Only group (M(SD)=448.2(352.9) minutes, range 0-1963 minutes, *NA*=22; Welch two-sample t-test: t=3.15, p=.002; **Figure 6.3**). Importantly, there was great variation in the total minutes spent listening to the audiobooks in both groups, ranging from 0 to 2500 minutes. If children listened to their books for the recommended time each week over the course of the intervention, they should have a total of 8 x 90=720 minutes total; this was higher than the mean for either group.

Figure 6.3: Children listened to audiobooks more in the Audiobooks+Scaffolding group.

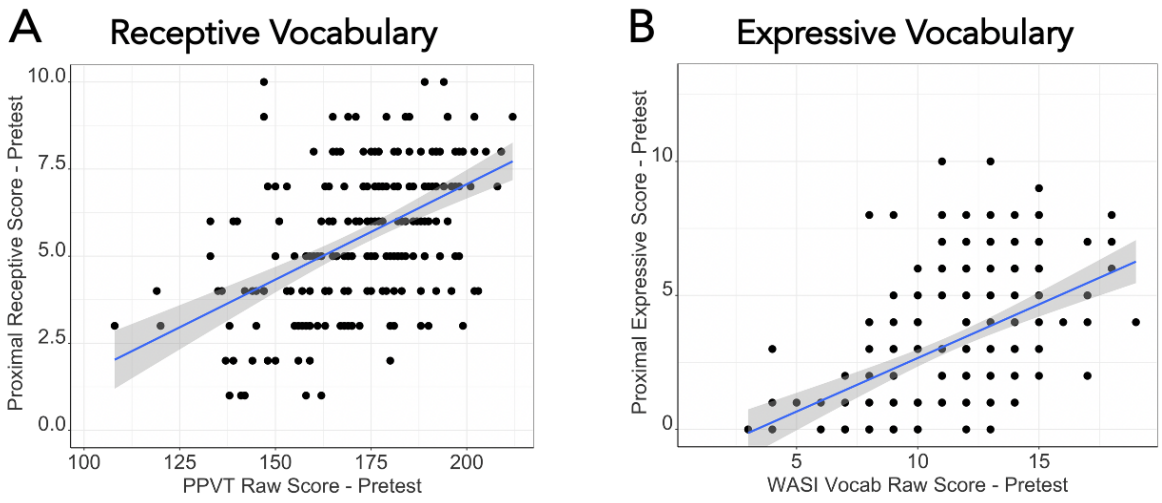


Boxplots show the total minutes spent listening to recommended books on the Learning Ally app during the study period (teal: Scaffolding+Audiobooks group; blue: Audiobooks-Only group). Each dot represents a participant. Red dot represents the group mean.

Proximal vocabulary measures were more sensitive than standardized measures to intervention effects.

To determine whether our proximal vocabulary measures captured meaningful variation in children's vocabulary skills, we compared participants' pretest scores for the two receptive vocabulary measures (PPVT raw score and final-book proximal receptive vocabulary score), and for the two expressive vocabulary measures (WASI vocabulary subtest raw score and final-l book proximal expressive vocabulary score). There was a positive correlation for the standard and proximal scores for both receptive (Pearson's correlation, r=.51, p<.001) and expressive (Pearson's correlation, r=.54, p<.001) vocabulary. Thus, we believe that the proximal measures are capturing some meaningful variation in children's vocabulary that is related to the standard measures they were based on.

Figure 6.4: Validity of proximal vocabulary assessments.



Scatterplots show raw scores for the standardized measures on the x-axis (PPVT; WASI) at pretest, and scores for the proximal measures on the y-axis, for ***(A)*** *receptive vocabulary and* ***(B)*** *expressive vocabulary.*

In preliminary analyses, we first aimed to measure intervention effects using the standardized vocabulary measures. We performed an exploratory analysis to measure the effects of age, reading fluency (Good Readers or Poor Readers based on DIBELS Passage Reading Fluency percentile), intervention group, and time (pretest vs. posttest) on receptive vocabulary as measured by the PPVT. There was a main effect of age and reading fluency (**Table 6.1, top**). Post-hoc Tukey tests revealed no gains for any of the groups, for Good Readers or Poor Readers (**Figure 6.5A**). Thus, by the standardized measure, there was no change in receptive vocabulary as a result of the intervention.

Next, we performed an exploratory analysis to measure the effects of age, reading fluency, group, and time on expressive vocabulary as measured by the WASI Vocabulary subtest. There was a main effect of age, reading fluency, and time, as well as an interaction between reading fluency, group, and time (**Table 6.1, bottom**). Surprisingly, post-hoc Tukey tests revealed significantly higher scores at posttest than pretest for Poor Readers in the Audiobooks+Scaffolding group, Good Readers in the Audiobooks-Only group, and both Good and Poor Readers in the Mindfulness group (**Figure 6.5B**).

Table 6.1: Preliminary effects of group and reading skills on standard vocabulary measures.

PPVT~age+PR*group*time+ (1|ParticipantID)						
Type III Analysis of Variance Table with Satterthwaite's method						
	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
age	**330.8**	**330.8**	**1**	**290.4**	**4.5**	**0.03***

PR	**1294.3**	**1294.3**	**1**	**296.8**	**17.8**	**<.001*** **
group	31.3	15.7	2	296.1	0.2	0.81
time	7.0	7.0	1	250.5	0.1	0.76
PR:group	196.3	98.2	2	296.4	1.3	0.26
PR:time	139.4	139.4	1	250.2	1.9	0.17
group:time	27.0	13.5	2	250.5	0.19	0.83
PR:group:time	5.7	2.9	2	250.1	0.04	0.96

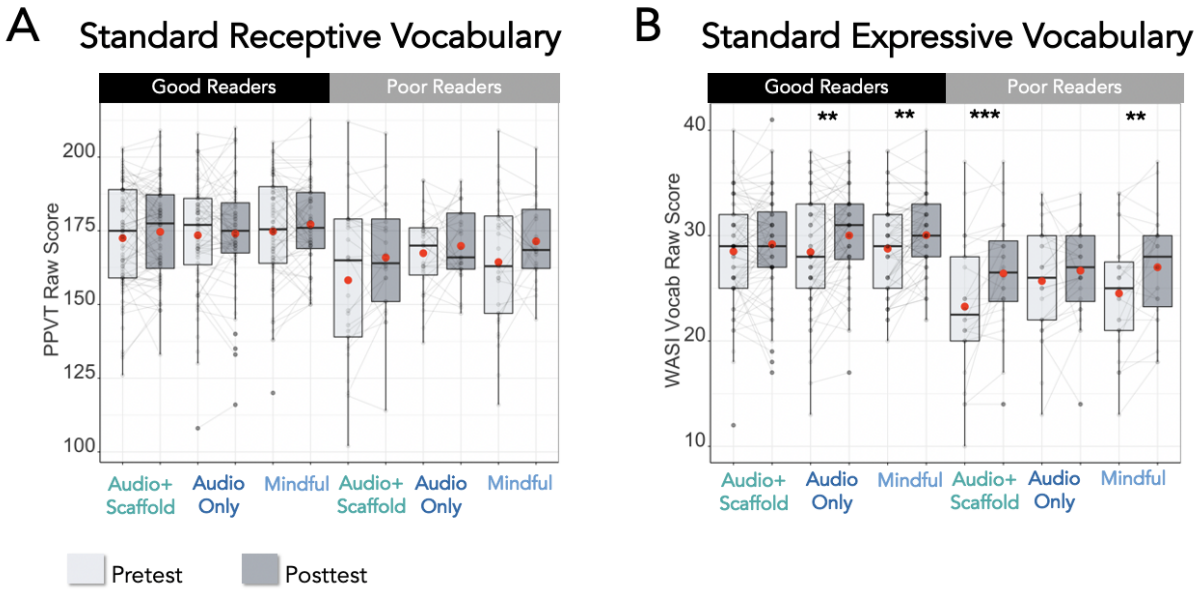
WASI~age+PR*group*time+ (1|ParticipantID)

Type III Analysis of Variance Table with Satterthwaite's method

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
age	**43.4**	**43.4**	**1**	**279.7**	**8.1**	**0.005****
PR	**199.0**	**199.0**	**1**	**284.3**	**37.0**	**<.001*** **
group	2.3	1.1	2	284.2	0.2	0.81
time	**92.7**	**92.7**	**1**	**251.3**	**17.2**	**<.001*** **
PR:group	3.3	1.7	2	284.0	0.3	0.73
PR:time	15.0	15.0	1	251.0	2.8	0.10
group:time	24.6	12.3	2	251.2	2.3	0.10
PR:group:time	**46.0**	**23.0**	**2**	**251.0**	**4.3**	**0.01***

PPVT = PPVT raw score; WASI = WASI Vocabulary subtest raw score; age = child age in years; PR = poor reader/good reader group based on reading fluency; group = intervention group (Audiobooks+Scaffolding, Audiobooks-Only, Mindfulness); time = pretest vs. posttest; Sum Sq = sum of squares; Mean Sq = mean sum of squares; NumDF = model degrees of freedom; DenDF = degrees of freedom associated with the model errors; F Value = F statistic; PR(>F) = p-value associated with the F-statistic.

Figure 6.5: Preliminary effects of group and reading skills on standard vocabulary measures.



*(A) Boxplots show PPVT raw scores at pretest and posttest, with individuals connected by light gray lines (red dot=mean). Good Readers are on the left, and Poor Readers are on the right. Participants are split by group. (B) WASI Vocabulary subtest raw scores at pretest and posttest. *p<.05, **p<.01, ***p<.001*

Next, we explored intervention effects using the proximal vocabulary measures. We performed an exploratory analysis to measure the effects of age, reading fluency, group, and time on receptive vocabulary as measured by the Proximal Receptive Vocabulary test. There was a main effect of age and time, as well as an interaction between reading fluency, group, and time (**Table 6.2, top**). Post-hoc Tukey tests revealed significantly higher scores at posttest for all groups among the Good Readers, but among the Poor Readers, only the Audiobooks+Scaffolding group showed significant gains at posttest (**Figure 6.6A**).

We also performed an exploratory analysis to measure the effects of age, reading fluency, group, and time on expressive vocabulary as measured by the Proximal Expressive Vocabulary test. There was a main effect of age, reading fluency, and time, and an interaction between group and time (**Table 6.2, bottom**). Post-hoc Tukey tests revealed significantly higher scores at posttest for Good Readers and Poor Readers in the Audiobooks+Scaffolding group, and only for Good Readers in the Audiobooks-Only group (**Figure 6.6B**).

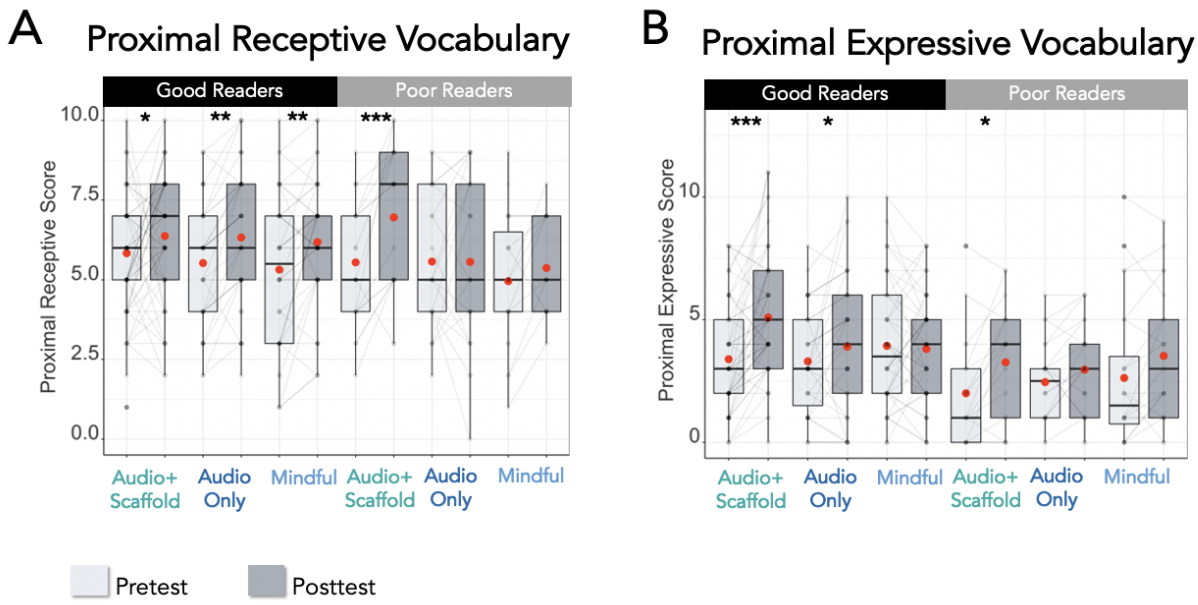
Table 6.2: Preliminary effects of group and reading skills on proximal vocabulary measures.

ProxRecVocab~age+PR*group*time+ (1|ParticipantID)						
Type III Analysis of Variance Table with Satterthwaite's method						
	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
age	**20.2**	**20.2**	**1**	**242.4**	**10.9**	**0.001****
PR	2.3	2.3	1	257.3	1.3	0.26
group	4.3	2.2	2	262.4	1.2	0.31
time	**41.2**	**41.2**	**1**	**220.3**	**22.3**	**<.001*****
PR:group	1.6	0.8	2	257.1	0.4	0.65
PR:time	0.02	0.02	1	211.8	0.01	0.92
group:time	2.8	1.4	2	218.6	0.8	0.47
PR:group:time	**12.2**	**6.1**	**2**	**211.4**	**3.3**	**0.04***
ProxExpVocab~age+PR*group*time+ (1|ParticipantID)						
Type III Analysis of Variance Table with Satterthwaite's method						

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
age	**24.4**	**24.4**	**1**	**244.9**	**11.5**	**<.001*** **
PR	**31.6**	**31.6**	**1**	**255.1**	**15.0**	**<.001*** **
group	5.9	2.9	2	259.8	1.4	0.25
time	**42.0**	**42.0**	**1**	**218.5**	**19.9**	**<.001*** **
PR:group	4.0	2.0	2	255.2	0.9	0.39
PR:time	0.03	0.03	1	212.8	0.02	0.90
group:time	**36.0**	**18.0**	**2**	**217.3**	**8.5**	**<.001*** **
PR:group:time	4.7	2.4	2	212.6	1.1	0.33

ProxRecVocab = Proximal Receptive Vocabulary test score; ProxExpVocab = Proximal Expressive Vocabulary test score; age = child age in years; PR = poor reader/good reader group; group = intervention group (Audiobooks+Scaffolding, Audiobooks-Only, Mindfulness); time = pretest vs. posttest; Sum Sq = sum of squares; Mean Sq = mean sum of squares; NumDF = model degrees of freedom; DenDF = degrees of freedom associated with the model errors; F Value = F statistic; PR(>F) = p-value associated with the F-statistic.

Figure 6.6: Preliminary effects of group and reading skills on proximal vocabulary measures.



*(A) Boxplots show Proximal Receptive Vocabulary scores at pretest and posttest, with individuals connected by light gray lines (red dot=mean). Good Readers are on the left, and Poor Readers are on the right. Participants are split by group. (B) Proximal Expressive Vocabulary scores at pretest and posttest. *p<.05, **p<.01, ***p<.001*

Discussion

We implemented a fully remote, randomized controlled trial (RCT) intervention to test whether listening to audiobooks improved children's language skills. This chapter presents preliminary analyses exploring the effects of the intervention on children's vocabulary. According to the standardized vocabulary measures, there was no evidence of intervention-specific improvements in vocabulary: no changes were seen in receptive vocabulary, and though there were pre-to-post changes in expressive vocabulary, they were in all three groups (Poor Readers in Audiobooks+Scaffolding, Good Readers in Audiobooks-Only, and both Good and Poor Readers in Mindfulness), and gains were therefore not specific to the audiobook groups. However, when using proximal vocabulary measures that included words children in the audiobooks groups

were exposed to in the books, pre-to-post gains were seen for poor readers only in the Audiobooks+Scaffolding group for both receptive and expressive vocabulary (good readers across all groups improved from pretest to posttest in receptive vocabulary, and good readers in both audiobook groups improved for expressive vocabulary).

First, these results provide some preliminary evidence that listening to audiobooks, when accompanied by scaffolded instructional support, may improve vocabulary skills for poor readers. Overall, good readers tended to improve vocabulary scores with time. Critically, for poor readers, it was only the Audiobooks+Scaffolding group that saw gains during the intervention. This is consistent with prior evidence that explicit and targeted instructional support may be especially important for struggling readers (Duff, 2019; Rupley et al., 2009), and also with prior work finding that just providing books over the summer does not improve reading scores (Kim, 2007). The scaffolding sessions explicitly taught and reviewed vocabulary words that appeared on the tests; thus, given that poor readers in the Audiobooks-Only group did not improve, it is possible that mere exposure from audiobooks was insufficient for vocabulary learning as measured by our assessments. However, because there was no condition with one-on-one scaffolding sessions without listening to audiobooks, it is unknown what role the audiobooks specifically played in these positive effects in the Audiobooks+Scaffolding group. Further work should explore whether vocabulary and scaffolding instruction paired with audiobooks, compared to with text-based books or without books at all, is specifically effective for poor readers.

These exploratory results also shed light on the importance of using carefully chosen outcome measures. When tested on the standardized vocabulary measures, children in the audiobook groups did not show specific gains, but when using measures that

included the words children encountered during the study period, the results differed. On one hand, this is very intuitive - why should children perform better on a standard vocabulary test if they did not encounter those words during the intervention period? Other studies examining the efficacy of vocabulary interventions have also found larger effects on proximal measures than standard measures of vocabulary (Apthorp et al., 2012; Elleman et al., 2009), and also for measures of comprehension (e.g., Vadasy et al., 2015). Standardized measures are fairly ubiquitous in educational research and can be important tools for comparing efficacy across studies; plus, educators typically look for improvements in standardized measures when implementing a research-supported approach in their classrooms (Solari et al., 2020). Norms from large samples help make these measures interpretable, and in general, standardized measures are rigorously developed. However, they may be insufficient for some purposes, and perhaps particularly for measuring intervention efficacy (e.g., see discussion in Elleman et al., 2009). Educational intervention studies are notorious for having small effect sizes and low transfer effects, and they often fail to replicate (Kim, 2019; Kraft, 2020; Lortie-Forgues & Inglis, 2019). Proximal measures, such as those included in our study, may be useful complementary measures in these cases. It is important to know whether a particular intervention can move the needle on standardized measures, but it may also be important to know whether they can move any needle at all.

Two important tradeoffs to creating individualized measures in the research setting are time and quality. It took over 10 research team members many months to create these tests, including selecting the words, generating the test items, and making the rubrics to score them. Each step was validated by independent experimenters, and there was lots of iteration along the way. Even with all these checks, some items had to later be

discarded. Because we were under extreme time pressure[27], we could not fully validate all the items on an independent sample prior to administration. Luckily scores on our proximal measures were correlated with standardized measures, which increased our confidence in using the measures – but if this had not been the case, the tests may not have been usable. Research on the impact of vocabulary on reading comprehension is of great interest and importance, but requires thoughtful development of vocabulary assessments (Pearson et al., 2007).

A second useful finding from this study is that children who received one-on-one scaffolded support during the intervention period spent more time listening to audiobooks than children who only received text-based reminders to listen to their books. Prior work has shown that adding scaffolded support to voluntary reading interventions can improve efficacy (Kim & White, 2008; White & Kim, 2008). Our work extends these findings, showing that scaffolding sessions can also increase intervention adherence for a remote, voluntary audiobook intervention. Another component of this voluntary intervention was choice in the books children read. Though children were constrained to a set of titles that were selected based on their listening comprehension skills, they could choose from a diverse set of options, including fiction and nonfiction books. Choice can motivate children to read more (Fisher & Frey, 2018; Guthrie et al., 2007), and a large body of research has also shown that children perform better on reading comprehension assessments when the material is interesting or familiar (Baldwin et al., 1985; Kendeou et al., 2003; Recht & Leslie, 1988; Shnayer, 1968). Thus, while not empirically tested in this study, it is likely that the element of student choice

[27] We needed to start administering these measures about a month after we decided to do this study in the first place.

impacted the efficacy of the study. Indeed, the majority of caregivers surveyed at the end of the study – in both audiobook groups – said that their child liked the recommended books "somewhat" or "a lot" (Ozernov-Palchik, Olson, et al., 2022; Chapter 5).

In considering these findings, it is important to note that this research study was conducted during a particular context: during the COVID-19 pandemic. The study began in summer 2020, and continued throughout the height of the pandemic as children and families navigated a mix of remote, hybrid, and in-person learning. Indeed, one motivation for conducting this study was to specifically try to support children from under-resourced backgrounds, as prior research on summer vacations suggested that students from lower-socioeconomic backgrounds tend to have greater learning loss over vacations than their higher-income peers (e.g., Cooper et al., 1996; however, recent research fails to replicate these effects: (von Hippel et al., 2018)), and that summer reading interventions may be particularly beneficial for low-income children (Kim & Quinn, 2013). Children with reading difficulties also lose ground over summer breaks (Christodoulou et al., 2017). There is substantial evidence that children experienced learning loss during the pandemic, with greater negative effects on children from disadvantaged environments (Donnelly & Patrinos, 2022; Engzell et al., 2021; Goldhaber et al., 2022; Khan & Ahmed, 2021). Furthermore, the pandemic greatly disrupted children's routines and social structures. Anecdotally, multiple caregivers in the Audiobooks+Scaffolding group commented that the scaffolding sessions were one of the only stable one-on-one relationships their child had outside the household during this time period. Given this context, it is therefore possible that this intervention approach may have had larger – or smaller – effects if conducted during a less tumultuous time.

Limitations

This study was complicated in many ways that we are in the process of untangling. We began in summer 2020 at the beginning of the COVID-19 pandemic, and we had to learn how to transition to remote research (see Ozernov-Palchik, Olson, et al., 2022 for discussion). As one might expect, there were some technical challenges and scheduling difficulties that led to certain tests being administered later than planned or missing altogether – like any intervention study, there were of course deviations from the 'ideal' plan. A benefit of administering the study remotely was that children were participating from all over the United States, from diverse socioeconomic backgrounds and experiencing widely varying schooling circumstances during the study period (learning in person, learning remotely, hybrid learning, vacation, etc.). The preliminary analyses in this chapter did not take into account myriad factors that may have impacted intervention adherence or outcomes. However, we do have substantial data on these factors and are planning to incorporate many of them into our preregistered confirmatory factor analysis, as well as additional exploratory analyses to address considerations that we did not foresee at the beginning of the study.

Conclusions

These preliminary results suggest that listening to audiobooks, particularly along with instructional support, may support vocabulary growth in struggling readers. Perhaps more importantly, they suggest that *how* vocabulary growth is measured, matters - measures that are created based on materials that children are explicitly exposed to may be more sensitive to intervention effects.

Acknowledgements

First, thank you to Ola Ozernov-Palchik, who will be my co-first author on the paper describing these results, as well as Xochitl Arechiga, who was particularly instrumental in this study. We are grateful for our partnership with *Learning Ally*, for helpful suggestions from James Kim and Tiffany Hogan, and for funding support from the Chan Zuckerberg Initiative for the Reach Every Reader project (https://www.gse.harvard.edu/reach-every-reader). Thank you to additional core project members: Yesi Camacho Torres, Hope Kentala, Natalie Gardino, Jeff Dieffenbach, Isaac Treves, Cindy Li, and Amanda Miller; undergraduate and high school research assistants: Jovita Solorio-Fielder, Tolu Asade, Cherry Wang, Sophia Angus, Bhuvna Murthy, Maycee McClure, Sehyr Khan, Hilary Zen, Sarah Abodalo, Elizabeth Carbonell, Shruti Das, Erika Leasher, Yoon Lim, Emmi Mills, Zoë Elizee, Camille Uldry, Shelby Laitipaya, Alexis Cho, Gabriella Aponte, Harley Yoder, Dana Osei, Niki Kim, Joy Bhattacharya, Emily Lin, Avni Ayer, Avery Dolins, and Abigail Cassidy; and testers and scorers: David Bates, Ross Weissman, Joohee Baik, June Okada, William Oliver, Harriet Richards, Kristen Wehara, Brooke Goldstein, and Ada Huang. Finally, we are grateful to the families for their time and participation.

References

Acheson, D. J., Wells, J. B., & MacDonald, M. C. (2008). New and updated tests of print exposure and reading abilities in college students. *Behavior Research Methods*, *40*(1), 278–289. https://doi.org/10.3758/BRM.40.1.278

Anderson, R. C., Wilson, P. T., & Fielding, L. G. (1988). Growth in Reading and How Children Spend Their Time Outside of School. *Reading Research Quarterly*, *23*(3), 285–303.

Apthorp, H., Randel, B., Cherasaro, T., Clark, T., McKeown, M., & Beck, I. (2012). Effects of a Supplemental Vocabulary Program on Word Knowledge and Passage Comprehension. *Journal of Research on Educational Effectiveness*, *5*(2), 160–188. https://doi.org/10.1080/19345747.2012.660240

Baldwin, R. S., Peleg-Bruckner, Z., & McClintock, A. H. (1985). Effects of Topic Interest and Prior Knowledge on Reading Comprehension. *Reading Research Quarterly*, *20*(4), 497–504. https://doi.org/10.2307/747856

Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing Words to Life: Robust Vocabulary Instruction*. Guilford Press.

Biber, D. (1991). *Variation Across Speech and Writing*. Cambridge University Press.

Brysbaert, M., & Biemiller, A. (2017). Test-based age-of-acquisition norms for 44 thousand English word meanings. *Behavior Research Methods*, *49*(4), 1520–1523. https://doi.org/10.3758/s13428-016-0811-4

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Catts, H. W., Adlof, S. M., & Weismer, S. E. (2006). Language Deficits in Poor Comprehenders: A Case for the Simple View of Reading. *Journal of Speech, Language, and Hearing Research*, *49*(2), 278–293. https://doi.org/10.1044/1092-4388(2006/023)

Christodoulou, J. A., Cyr, A., Murtagh, J., Chang, P., Lin, J., Guarino, A. J., Hook, P., & Gabrieli, J. D. E. (2017). Impact of Intensive Summer Reading Intervention for Children With Reading Disabilities and Difficulties in Early Elementary School. *Journal of Learning Disabilities*, *50*(2), 115–127. https://doi.org/10.1177/0022219415617163

Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The Effects of Summer Vacation on Achievement Test Scores: A Narrative and Meta-Analytic Review. *Review of Educational Research*, *66*(3), 227–268. https://doi.org/10.3102/00346543066003227

Donnelly, R., & Patrinos, H. A. (2022). Learning loss during Covid-19: An early systematic review. *PROSPECTS*, *51*(4), 601–609. https://doi.org/10.1007/s11125-021-09582-6

Duff, D. (2019). The Effect of Vocabulary Intervention on Text Comprehension: Who Benefits? *Language, Speech, and Hearing Services in Schools*, *50*(4), 562–578. https://doi.org/10.1044/2019_LSHSS-VOIA-18-0001

Duff, D., Tomblin, J. B., & Catts, H. (2015). The Influence of Reading on Vocabulary Growth: A Case for a Matthew Effect. *Journal of Speech, Language, and Hearing Research*, *58*(3), 853–864. https://doi.org/10.1044/2015_JSLHR-L-13-0310

Dunn, D. M. (2018). *Peabody Picture Vocabulary Test* (5th ed.). NCS Pearson.

Elleman, A. M., Lindo, E. J., Morphy, P., & Compton, D. L. (2009). The Impact of Vocabulary Instruction on Passage-Level Comprehension of School-Age Children: A Meta-Analysis. *Journal of Research on Educational Effectiveness*, *2*(1), 1–44. https://doi.org/10.1080/19345740802539200

Engzell, P., Frey, A., & Verhagen, M. D. (2021). Learning loss due to school closures during the COVID-19 pandemic. *Proceedings of the National Academy of Sciences*, *118*(17), e2022376118. https://doi.org/10.1073/pnas.2022376118

Fisher, D., & Frey, N. (2018). Raise Reading Volume Through Access, Choice, Discussion, and Book Talks. *The Reading Teacher*, *72*(1), 89–97. https://doi.org/10.1002/trtr.1691

Goldhaber, D., Kane, T. J., McEachin, A., Morton, E., Patterson, T., & Staiger, D. O. (2022). *The Consequences of Remote and Hybrid Instruction During the Pandemic* (Working Paper No. 30010). National Bureau of Economic Research. https://doi.org/10.3386/w30010

Good, R. H., & Kaminski, R. A. (Eds.). (2002). *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.). Institute for the Development of Educational Achievement. http://dibels.uoregon.edu/

Goodwin, A. P. (2016). Effectiveness of word solving: Integrating morphological problem-solving within comprehension instruction for middle school students. *Reading and Writing*, *29*(1), 91–116. https://doi.org/10.1007/s11145-015-9581-0

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading and reading disability. *Remedial and Special Education*, 6–10.

Guerreiro, M., Barker, E., & Johnson, J. (2022). Measuring Student Reading Comprehension Performance: Considerations of Accuracy, Equity, and Engagement by Embedding Comprehension Items within Reading Passages. *Practical Assessment, Research, and Evaluation*, *27*(1). https://doi.org/10.7275/ch8r-tx33

Guthrie, J. T., Hoa, A. L. W., Wigfield, A., Tonks, S. M., Humenick, N. M., & Littles, E. (2007). Reading motivation and reading comprehension growth in the later elementary years. *Contemporary Educational Psychology*, *32*(3), 282–313. https://doi.org/10.1016/j.cedpsych.2006.05.004

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, *2*(2), 127–160. https://doi.org/10.1007/BF00401799

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Brief Intelligence Test* (2nd ed.). American Guidance Service.

Kendeou, P., Rapp, D. N., & van den Broek, P. (2003). The influence of reader's prior knowledge on text comprehension and learning from text. In R. Nata (Ed.), *Progress in Education* (Vol. 13, pp. 189–209). Nova Science Publishers, Inc.

Kendeou, P., van den Broek, P., White, M. J., & Lynch, J. S. (2009). Predicting reading comprehension in early elementary school: The independent contributions of oral language and decoding skills. *Journal of Educational Psychology*, *101*(4), 765–778. https://doi.org/10.1037/a0015956

Khan, M. J., & Ahmed, J. (2021). Child education in the time of pandemic: Learning loss and dropout. *Children and Youth Services Review*, *127*, 106065. https://doi.org/10.1016/j.childyouth.2021.106065

Kim, J. S. (2007). The effects of a voluntary summer reading intervention on reading activities and reading achievement. *Journal of Educational Psychology*, *99*, 505–515. https://doi.org/10.1037/0022-0663.99.3.505

Kim, J. S. (2019). Making Every Study Count: Learning From Replication Failure to Improve Intervention Research. *Educational Researcher*, *48*(9), 599–607. https://doi.org/10.3102/0013189X19891428

Kim, J. S., & Quinn, D. M. (2013). The Effects of Summer Reading on Low-Income Children's Literacy Achievement From Kindergarten to Grade 8: A Meta-Analysis of Classroom and Home Interventions. *Review of Educational Research*, *83*(3), 386–431. https://doi.org/10.3102/0034654313483906

Kim, J. S., & White, T. G. (2008). Scaffolding Voluntary Summer Reading for Children in Grades 3 to 5: An Experimental Study. *Scientific Studies of Reading*, *12*(1), 1–23. https://doi.org/10.1080/10888430701746849

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

Language and Reading Research Consortium. (n.d.). *Language and Reading Research Consortium*. Retrieved March 3, 2023, from https://larrc.ehe.osu.edu/curriculum/downloads/

Language and Reading Research Consortium. (2016). Use of the Curriculum Research Framework (CRF) for Developing a Reading-Comprehension Curricular Supplement for the Primary Grades. *The Elementary School Journal*, *116*(3), 459–486. https://doi.org/10.1086/684827

Language and Reading Research Consortium, Arthur, A. M., & Davis, D. L. (2016). A Pilot Study of the Impact of Double-Dose Robust Vocabulary Instruction on Children's Vocabulary Growth. *Journal of Research on Educational Effectiveness*, *9*(2), 173–200. https://doi.org/10.1080/19345747.2015.1126875

Language and Reading Research Consortium, Pratt, A., & Logan, J. (2014). Improving Language-Focused Comprehension Instruction in Primary-Grade Classrooms: Impacts of the Let's Know! Experimental Curriculum. *Educational Psychology Review*, *26*(3), 357–377. https://doi.org/10.1007/s10648-014-9275-1

Lortie-Forgues, H., & Inglis, M. (2019). Rigorous Large-Scale Educational RCTs Are Often Uninformative: Should We Be Concerned? *Educational Researcher*, *48*(3), 158–166. https://doi.org/10.3102/0013189X19832850

Melekoğlu, M. A., & Wilkerson, K. L. (2012). Motivation to Read: How Does It Change for Struggling Readers with and without Disabilities? *International Journal of Instruction*, *6*(1), Article 1.

Montag, J. L., Jones, M. N., & Smith, L. B. (2015). The Words Children Hear: Picture Books and the Statistics for Language Learning. *Psychological Science*, *26*(9), 1489–1496. https://doi.org/10.1177/0956797615594361

Montag, J. L., & MacDonald, M. C. (2015). Text exposure predicts spoken production of complex sentences in 8- and 12-year-old children and adults. *Journal of Experimental Psychology: General*, *144*, 447–468. https://doi.org/10.1037/xge0000054

NAEP Reading: Reading Highlights 2022. (n.d.). Retrieved March 9, 2023, from https://www.nationsreportcard.gov/highlights/reading/2022/

National Reading Panel (U.S.). (2000). *Teaching Children to Read: An Evidence-based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction : Reports of the Subgroups*. National Institute of Child Health and Human Development, National Institutes of Health.

Ozernov-Palchik, O., Olson, H. A., Arechiga, X. M., Kentala, H., Solorio-Fielder, J. L., Wang, K. L., Torres, Y. C., Gardino, N. D., Dieffenbach, J. R., & Gabrieli, J. D. E. (2022). Implementing Remote Developmental Research: A Case Study of a Randomized Controlled Trial Language Intervention During COVID-19. *Frontiers in Psychology*, *12*, 6163. https://doi.org/10.3389/fpsyg.2021.734375

Paris, S. G., & Stahl, S. A. (2005). *Children's Reading Comprehension and Assessment*. Routledge.

Pearson, P. D., Hiebert, E. H., & Kamil, M. L. (2007). Vocabulary assessment: What we know and what we need to learn. *Reading Research Quarterly*, *42*(2), 282–296. https://doi.org/10.1598/RRQ.42.2.4

Recht, D. R., & Leslie, L. (1988). Effect of prior knowledge on good and poor readers' memory of text. *Journal of Educational Psychology*, *80*, 16–20. https://doi.org/10.1037/0022-0663.80.1.16

Rupley, W. H., Blair, T. R., & Nichols, W. D. (2009). Effective Reading Instruction for Struggling Readers: The Role of Direct/Explicit Teaching. *Reading & Writing Quarterly*, *25*(2–3), 125–138. https://doi.org/10.1080/10573560802683523

Scarborough, H. S. (2001). Connecting early language and literacy to later reading (dis) abilities: Evidence, theory, and practice. In S. Neuman & D. Dickinson (Eds.), *Handbook for research in early literacy* (pp. 97–110). Guilford Press.

Sénéchal, M. (1997). The differential effect of storybook reading on preschoolers' acquisition of expressive and receptive vocabulary. *Journal of Child Language*, *24*(1), 123–138. https://doi.org/10.1017/S0305000996003005

Shnayer, S. W. (1968). *Some Relationships between Reading Interest and Reading Comprehension*. https://eric.ed.gov/?id=ED022633

Singh, A., & Alexander, P. A. (2022). Audiobooks, Print, and Comprehension: What We Know and What We Need to Know. *Educational Psychology Review*, *34*(2), 677–715. https://doi.org/10.1007/s10648-021-09653-2

Solari, E. J., Terry, N. P., Gaab, N., Hogan, T. P., Nelson, N. J., Pentimonti, J. M., Petscher, Y., & Sayko, S. (2020). Translational Science: A Road Map for the Science of Reading. *Reading Research Quarterly*, *55*(S1), S347–S360. https://doi.org/10.1002/rrq.357

Stahl, S. A., & Fairbanks, M. M. (1986). The Effects of Vocabulary Instruction: A Model-Based Meta-Analysis. *Review of Educational Research*, *56*(1), 72–110. https://doi.org/10.3102/00346543056001072

Tilstra, J., McMaster, K., Van den Broek, P., Kendeou, P., & Rapp, D. (2009). Simple but complex: Components of the simple view of reading across grade levels. *Journal of Research in Reading*, *32*(4), 383–401. https://doi.org/10.1111/j.1467-9817.2009.01401.x

Troyer, M., Kim, J. S., Hale, E., Wantchekon, K. A., & Armstrong, C. (2019). Relations among intrinsic and extrinsic reading motivation, reading amount, and comprehension: A conceptual replication. *Reading and Writing*, *32*(5), 1197–1218. https://doi.org/10.1007/s11145-018-9907-9

Vadasy, P. F., Sanders, E. A., & Logan Herrera, B. (2015). Efficacy of Rich Vocabulary Instruction in Fourth- and Fifth-Grade Classrooms. *Journal of Research on Educational Effectiveness*, *8*(3), 325–365. https://doi.org/10.1080/19345747.2014.933495

van Bergen, E., Snowling, M. J., de Zeeuw, E. L., van Beijsterveldt, C. E. M., Dolan, C. V., & Boomsma, D. I. (2018). Why do children read more? The influence of reading ability on voluntary reading practices. *Journal of Child Psychology and Psychiatry*, *59*(11), 1205–1214. https://doi.org/10.1111/jcpp.12910

von Hippel, P. T., Workman, J., & Downey, D. B. (2018). Inequality in Reading and Math Skills Forms Mainly before Kindergarten: A Replication, and Partial Correction, of "Are Schools the Great Equalizer?" *Sociology of Education*, *91*(4), 323–357. https://doi.org/10.1177/0038040718801760

Vygotsky, L. S., & Cole, M. (1978). *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press.

Wasik, B. A., Hindman, A. H., & Snell, E. K. (2016). Book reading and vocabulary development: A systematic review. *Early Childhood Research Quarterly*, *37*, 39–57. https://doi.org/10.1016/j.ecresq.2016.04.003

Wechsler, D. (2011). *Wechsler Abbreviated Scale of Intelligence–Second Edition (WASI-II)*. NCS Pearson.

White, T. G., & Kim, J. S. (2008). Teacher and Parent Scaffolding of Voluntary Summer Reading. *The Reading Teacher*, *62*(2), 116–125. https://doi.org/10.1598/RT.62.2.3

Wood, S. G., Moxley, J. H., Tighe, E. L., & Wagner, R. K. (2018). Does Use of Text-to-Speech and Related Read-Aloud Tools Improve Reading Comprehension for Students With Reading Disabilities? A Meta-Analysis. *Journal of Learning Disabilities*, *51*(1), 73–84. https://doi.org/10.1177/0022219416688170

Wright, T. S., & Cervetti, G. N. (2017). A Systematic Review of the Research on Vocabulary Instruction That Impacts Text Comprehension. *Reading Research Quarterly*, *52*(2), 203–226. https://doi.org/10.1002/rrq.163

Supplementary

SUPPLEMENTARY TABLE 1: Book tracks.

Track 1	Track 2	Track 3
	Track Assignment: 3rd Grade: ● CELF<10 4th Grade: ● CELF<7 & PPVT<100	Track Assignment: 3rd Grade: ● CELF≥10 4th Grade: ● CELF<7 & PPVT≥100 ● CELF≥7 & PPVT<100 ● CELF≥7 & PPVT≥100
Primary: 1. *Hank Zipzer*, by Henry Winkler and Lin Oliver (FIRST) 2. *Memphis, Martin, and the Mountaintop : The Sanitation Strike of 1968*, by	Primary: 1. *Mr. Klutz Is Nuts*, by Dan Gutman (FIRST) 2. *Amina's Voice*, by Hena Khan 3. *Who Was Galileo?*, by Patricia Brennan Demuth 4. *Frindle*, by Andrew Clements	Primary: 1. *Frindle*, by Andrew Clements (FIRST) 2. *Schomburg: The Man Who Built A Library*, by Carole Boston Weatherford 3. *The Bad Beginning*, by Lemony Snicket

Alice Faye Duncan 3. *The Boy Who Invented TV : The Story of Philo Farnsworth*, by Kathleen Krull 4. *The Chocolate Touch*, by Patrick Skene Catling 5. *I Survived The Attack of the Grizzlies, 1967*, by Lauren Tarshis 6. *Who Was Maya Angelou?*, by Ellen Labrecque (LAST) <u>Additional titles</u>: ● *How To Eat Fried Worms*, by Thomas Rockwell ● *The One And Only Ivan*, by Katherine Applegate ● *The Mystery of the Missing Cat*, by Gertrude Chandler Warner ● *Tales Of A Fourth Grade Nothing*, by Judy Blume	5. *Thirty Minutes Over Oregon: A Japanese Pilot's World War II Story*, by Marc Tyler Nobleman (LAST) <u>Additional titles</u>: ● *The Lemonade War*, by Jacqueline Davies ● *We Are The Ship : The Story of Negro League Baseball*, by Kadir Nelson ● *Who Was Maya Angelou?*, by Ellen Labrecque ● *The Chocolate Touch*, by Patrick Skene Catling ● *Memphis, Martin, and the Mountaintop : The Sanitation Strike of 1968*, by Alice Faye Duncan ● *Tales Of A Fourth Grade Nothing*, by Judy Blume	4. *Puppies Dogs and Blue Northers : Reflections on Being Raised by a Pack of Sled Dogs*, by Gary Paulsen 5. *Crenshaw*, by Katherine Applegate (LAST) <u>Additional titles</u>: ● *Chasing Space Young Readers' Edition*, by Leland D. Melvin ● *Thirty Minutes Over Oregon: A Japanese Pilot's World War II Story*, by Marc Tyler Nobleman ● *We Are The Ship : The Story of Negro League Baseball*, by Kadir Nelson ● *Amina's Voice*, by Hena Khan ● *The Reptile Room*, by Lemony Snicket ● *Bob*, by Wendy Mass ● *Young Captain Nemo*, by Jason Henderson ● *Lifeboat 12*, by Susan Hood

Participants in the audiobook groups were assigned to a book track based on their grade, listening comprehension (CELF standard score), and receptive vocabulary (PPVT standard score). These assignments could change based on children's experiences with

the first book. Typically, participants read the designated first book and last book; however, if children reported having read the book previously, they were tested on and read another book instead. These book tracks were created by the Learning Ally team for our study to include some grade level books and some above-grade-level books, balance fiction and nonfiction, and include diverse authors and topics.

SUPPLEMENTARY TABLE 2: Audiobooks+Scaffolding lesson plan summary.

Week	Session 1	Session 2
1	Sequencing	Comprehension Monitoring
2	Text Mapping Noun Phrases	Retelling
3	Main Character	Predicting
4	Text Mapping Verb Phrases	Author's Purpose
5	Text Mapping Prefixes	Character's Goals
6	Reporting	Prediction
7	Text Mapping - Comparing Characters	Sequencing
8	Alternate Outcomes	Alternate Outcomes

Lesson plans were adapted from the Language and Reading Research Consortium materials.

"The story itself, the true story, is the one that the audience members create in their minds, guided and shaped by my text, but then transformed, elucidated, expanded, edited, and clarified by their own experience, their own desires, their own hopes and fears."

— Orson Scott Card, ***Ender's Game***

Chapter 7 : Discussion

Studying the neural basis of language development feels akin to studying the building blocks of our human experience. Languages come in many forms and in multiple modalities – yet the underlying neural architecture is remarkably consistent. We use language to communicate with others, and sometimes we use language to think and process the world for ourselves. We teach our children to read and write, so that language can stand on its own and communicate our thoughts, knowledge, and ideas even when we are not present. Language is both insufficient to express the fullness of our communicative goals, and also independently generative of new ideas that supersede the thoughts of the original speaker. How do we come to embrace language as core to who we are and how we interact with the world?

Language is subserved by a specific network in the brain, no matter what language someone speaks (e.g., Honey et al., 2012; Malik-Moraleda et al., 2022). Language develops whether it is heard, seen, or felt (e.g., Bedny et al., 2011; Fedorenko et al., 2010; MacSweeney et al., 2008; Neville et al., 1998; Obretenova et al., 2010; Scott et al., 2017) – it is a beautiful example of the interplay between experience and endogenous constraints on development. In this small body of work, I examined ways

in which language processing, at the level of both brain and behavior, can be influenced by these factors at various points in development.

In so doing, I came to appreciate a few particular challenges inherent in studying language, and the brain, and development. First, language – though we study it as an amodal, flexible cognitive tool, it is impossible to completely extricate the core of 'language' from the context in which it is learned and used. This is especially consequential when trying to study the neural basis of language, as functional neuroimaging relies on either deliberately controlled study design to isolate a cognitive construct of interest, which imparts researchers' biases and constraints, or deliberately uncontrolled stimuli that vary in such a way as to elicit reliable patterns of activity in the brain, which may not allow for a construct like 'language' to be extracted from the signal at all. And then we come to development – what is 'language' in a two-year-old? In an eight-year-old? When the cognitive construct itself is changing over time, and the brain is changing too, how do we disentangle the effects of endogenous and exogenous factors to isolate a developmental trajectory?

The preceding studies do not answer these questions, but they hopefully add a couple useful approaches to our collective toolbox. Below I summarize the main findings of these studies, what I believe to be the main takeaways of this work, and where I think we should go next.

A summary of the preceding chapters

This work underscored two main themes: first, innovating techniques for measuring language activation in the brain in difficult-to-reach developmental populations, and

second, conceptualizing what it means to measure language competence while accounting for variability in children's interests and experiences.

Chapters 2-3 described a novel experimental approach to studying language in the brains of toddlers (and adults). The guiding ethos of this project was engagement – what would pull toddlers in, enough to keep them in an MRI scanner? Many other developmental researchers had already realized the benefits of using naturalistic stimuli in neuroimaging experiments with infants (e.g., Kosakowski et al., 2022; Yates et al., 2022) and children (Cantlon, 2020; Cantlon & Li, 2013; Kamps et al., 2022; Redcay & Moraczewski, 2020; Richardson et al., 2018; Vanderwal et al., 2019), but toddlers are a new frontier for awake fMRI. So we turned to the experts, and selected clips from a television show with decades of educational programming directed at our target age group: *Sesame Street* (following the intuition of Cantlon & Li, 2013; Emerson & Cantlon, 2012, who also used *Sesame Street* in fMRI). Critically, we needed to ensure that toddlers not only attended to the comprehensible language condition, but also to a control condition, so we embedded backwards speech into the video clips by reversing the audio of one character at a time. First, we validated that our approach of embedding forward and backward speech within naturalistic video clips 'worked' by testing it on adults – and indeed, language regions in adults responded more to forward than backward speech[28]. Along the way, we found that language regions did not respond differently to the dialogue videos than the monologue ones, but that some theory of mind regions, and some right hemisphere homologues of language regions, did respond more to dialogue. These results will be integral for grounding interpretations of data from toddlers. Next, we tested the task behaviorally, confirming

[28] Phew.

that toddlers will attend to all the task conditions. Finally, we began scanning awake toddlers, finding preliminary evidence that we can evoke - and measure - language-related activity in canonical language regions.

Chapter 4 took this ethos of engagement to the next level. Inspired by the strength of personal interests in the everyday lives of many children – and, in particular, many autistic children (Klin et al., 2007) – we sought to determine whether personal interest modulated activation in language regions of the brain during language comprehension. Autistic and neurotypical children were scanned while listening to short narratives about their absolute favorite topic, and also while listening to short narratives about nature (no one's favorite topic in this sample). Canonical language regions, as well as cortical and subcortical regions associated with reward and salience, robustly responded more to the personalized narratives about each child's interest in both groups.

Finally, Chapters 5-6 pivoted to a study of behavior rather than the brain. During the pandemic, we conducted a remote, randomized controlled trial intervention to test the effects of listening to audiobooks on children's language skills. We learned how to adapt our experimental approaches to a new setting, including the benefits and challenges of recruiting a diverse geographic, skill-level, and socioeconomic sample. Preliminary analyses suggested that audiobooks paired with instructional support may improve struggling readers' vocabulary skills. But it mattered how we measured those skills – specifically, we saw evidence of this change only when using proximal measures that included words in the audiobooks.

These results contribute to iterative scientific progress in their own right: language regions do not seem to differentiate between dialogue and monologue in adult brains, language regions may be sensitive to how interesting language is to the individual child, and listening to audiobooks paired with instructional support may impact vocabulary for struggling readers. However, they also spark questions about thornier issues. How do we go about studying language development, especially in the brain, and how do our methodological choices impact the inferences we can make from the data we collect?

Impact of endogenous and exogenous factors on language development

One goal of this dissertation was to explore some of the myriad factors that impact language development in children – including endogenous factors, like language difficulties and personal interests, as well as exogenous factors, like the social context of language exposure and the words in books children read – to see whether these factors impact brain function during language processing.

Language in a social context

Language and social processing are fundamentally intertwined – yet despite decades of research on the cortical system engaged in language processing, surprisingly few studies have tested the neural mechanisms underlying comprehension of conversation (though see Bašnáková et al., 2014; Feng et al., 2017; Jang et al., 2013, for a few examples). Using audiovisual conversational stimuli, Chapter 2 provided additional evidence that social context is not processed by canonical language regions in adults: language regions did not differentiate between monologue and dialogue speech. This

is consistent with previous evidence that language and social processing are functionally distinct in the adult brain (e.g., (Paunov et al., 2019, 2022)). However, Chapter 3 provides an approach for empirically testing an open question: whether this distinction is also present in younger children. Behavioral evidence has suggested that social context is important for language learning during development (Golinkoff et al., 2015; Hoff, 2006; Kuhl, 2007), and that early socio-linguistic interactions can impact language skills (Ferjan Ramírez et al., 2020; Hirsh-Pasek et al., 2015; Romeo et al., 2018; Rowe, 2008). Indeed, we also found preliminary evidence in Chapter 6 that struggling readers' vocabulary only improved when audiobooks were paired with one-on-one instructional sessions (i.e., *social interaction*). Furthermore, some have argued that language evolved to serve a social function (e.g., Seyfarth & Cheney, 2014). While the preliminary results from toddlers in Chapter 3 are too early to interpret, it is plausible that the neural underpinnings of language processing are not completely distinct from processing social context at this stage of development.

Language content in development

Another factor we examined in these studies was how the content of language can impact language learning and function in children. In Chapter 6, we found preliminary evidence that listening to audiobooks paired with instructional support may be an effective way to increase children's vocabulary. Prior work has suggested that reading books exposes children to more complex vocabulary and sentence structures than everyday speech (Biber, 1991; Montag et al., 2015; Montag & MacDonald, 2015). While preliminary, our work points to *content*, rather than the act of reading per se, that may play a role in vocabulary learning – thus, directly comparing these modalities is an important direction for future work. In children, content can also impact brain

activation within language regions, as we observed in Chapter 4: language regions responded more to personally-interesting content than non-personally-interesting narratives. Understanding how, and why, is another important future direction for developmental language research.

Overall, a takeaway from this work is that both context and content play a role in children's language processing, and thus may be particularly important to take into account when studying language development.

Measuring language development

In the summary section above, I linked my projects together by alluding to a so-called 'ethos of engagement' that pervaded the study design. This, too, links the two themes I articulated in the introduction – it is by leaning in to variability in the exogenous and endogenous factors that contribute to language that we were able to innovate new techniques to study language in the developing brain, specifically by *engaging* the target population. An overarching takeaway from this work is that it matters *how* we study language, especially in young populations, both in terms of who ends up in our research and how we interpret subsequent results.

How do we tailor stimuli for developmental populations?

Most neuroimaging studies assume that generic language is sufficient to elicit a reliable 'language' response in the brain. This is both logical and prudent. First, because it tends to work in adults, and second, because it allows for lower-level experimental control (e.g., Fedorenko et al., 2010). However, in these studies, I pointed to cases in which generic language stimuli *do not* work in developmental populations, for

methodological reasons (e.g., getting toddlers to attend to a task), for sensitivity reasons (e.g., detecting whether an intervention had an effect on vocabulary), and for interpretation reasons (e.g., inferring the extent to which certain regions are involved in language processing).

These cases point to a deeper question: what does it mean to have experimental control, specifically for language? In Chapter 4, we suggest that when studying language in the brain, it is perhaps just as important to consider controlling for a higher-level confound like personal interest as it is to control for acoustic properties of speech. Indeed, it may even be *necessary* to compromise on one level of control in order to achieve control at another level. In Chapters 2-3, for instance, we opted for backwards speech as the control condition rather than a 'better matched' acoustically degraded speech because we thought that the backwards speech worked better for maintaining the impression of two characters talking to each other. In Chapter 4, we again compromised on lower-level control by not matching the vocabulary level between the personally-interesting and neutral narratives, because a key component of veridically conveying an interest is using the correct terminology. By an 'objective' measure of vocabulary, a word like 'Flareon' may be fairly complex – yet for a child that plays Pokémon every day, this word is quite familiar[29].

What, then, does it mean to tailor stimuli to measure language development? Do we mean by topic? By social mode of language delivery? By exposure to specific words? Just as language is infinitely generative, there are myriad ways to tailor language

[29] Flareon is a mammalian fire-type Pokémon that is pretty cute and totally awesome.

stimuli. Deciding how – and whether – stimuli should be tailored will depend on the goal of the study.

Tailoring stimuli for methodological reasons is common in developmental psychology, though it is not without controversy. One example of adapting stimuli for the target population is the use of simplified, colorful puppets in infant behavioral studies in order to minimize unintentional confounds and distractions (e.g., Kominsky et al., 2022). While some have maintained that this adaptation reduces ecological validity (Packer & Moreno-Dulcey, 2022), the arguably bigger concern is that these kinds of studies would not work at all *without* tailoring the stimuli. This was similarly the concern in the toddler fMRI study (Chapter 3). Rather than a purely auditory task, we anticipated that toddlers would be more compliant when watching videos in the scanner. Indeed, in older children undergoing resting state scans, data quality is better when children watch movies than when they are told to stare at a fixation cross (Frew et al., 2022). But which videos to use? In order to impose some experimental control, we needed sufficient source material that included certain types of scenes: monologue and dialogue, without other characters present, with some variability in background and content. We also decided to use puppets, as the rigid mouth movements were easier to align with the backwards speech, since young children are sensitive to mismatches in auditory and visual speech cues (Gogate & Bahrick, 1998; Lewkowicz & Flom, 2014). Tailoring stimuli for methodological reasons requires an intentional and iterative process – including, importantly, testing to see if the stimuli actually hold the attention of the target population, as we did in Chapter 3.

Another reason to tailor experimental stimuli, as demonstrated in Chapter 6, is to more closely align with individual differences in experience, and thereby increase the

sensitivity of the measure. After an 8-week intervention, for instance, why would we expect children to perform better on a vocabulary test if they did not encounter any of the words on it? Many educational intervention studies have found that proximal measures are more sensitive to intervention effects than standard measures (e.g., Apthorp et al., 2012; Elleman et al., 2009; Vadasy et al., 2015), and indeed, the proximal vocabulary measures seem to be more sensitive in our study (Chapter 6). However, it is important to ensure that a new measure actually captures the underlying cognitive construct of interest. As a toy example, imagine an intervention study that aims to improve reading comprehension by training children to solve multiplication problems, based on a correlation between reading and math skills. A proximal measure of the intervention's efficacy might be how well children solve multiplication problems, whereas a distal measure might be a standardized reading assessment. If the child improves on the "proximal" measure of multiplication skills, and not on the more "distal" measure of reading comprehension, we obviously should not infer that the reading measure simply was not *sensitive* enough to detect an effect. Tailoring experimental stimuli to increase sensitivity requires ensuring that the new measure maintains construct validity. In Chapter 6, we designed our proximal vocabulary tests based on the standardized assessments, and also checked to make sure the scores on both tests were correlated at pretest.

Finally, tailoring stimuli can be important when measuring a complex cognitive construct like language in order to ensure that measures are not confounded by other factors. For example, reading comprehension can be impacted by how familiar or interesting the material is to a child (Baldwin et al., 1985; Kendeou et al., 2003; Recht & Leslie, 1988; Shnayer, 1968). In Chapter 4, we found higher activation in language network – as well as more extensive activation – for narratives written about a child's

personal interest, compared to non-personalized narratives about nature. When studying individual differences in language activation, one would not want to confuse language-evoked responses with individual differences in levels of interest. Particularly if one anticipates widely varying levels of interest in the stimuli, it may be worth tailoring those stimuli to topics that will engage each child. In so doing, it is important to maintain a balance between imposing experimental control when possible – e.g., by using the same speaker for language clips – but also to also capture the spirit of the interest – e.g., by using interest-specific terminology.

These examples are just a few ways in which language stimuli can be tailored to different developmental populations, and the preceding studies show the feasibility and promise of these approaches. A valid concern, though, is how to determine whether one has maintained experimental control when making these adaptations.

How do we know if we have tailored stimuli well?

In addition to figuring out how to tailor the stimuli, it is equally important (and time consuming) to make sure it is done properly - as the consequences of poorly adjusting stimuli could be meaningful. However, trying to determine whether we have appropriately tailored our research approach can quickly become circular. For example, previous research suggested that language network may be atypical or underactive in autism (Mody & Belliveau, 2013). We hypothesized, however, that heterogeneity in the brain's response to language could plausibly be an effect of how interesting and engaging the stimuli were to each individual, particularly because autistic individuals often have highly restricted special interests (Klin et al., 2007). Thus, we personalized language stimuli to the individual based on their interests, and found that, indeed,

there is higher activation and also more widespread activation for personally-interesting than neutral language. But herein lies the puzzle – what is the 'language' response? How do we know whether the non-personalized stimuli were 'missing' areas, or whether the personalized stimuli were recruiting additional regions?

There is no simple answer to this puzzle, but there are a few approaches that might increase our confidence in using 'less controlled' experimental stimuli. First, certain analytical approaches might be more robust to lower-level differences in the stimuli. Subject-specific functional localization (Fedorenko et al., 2010; Kanwisher et al., 1997; Saxe et al., 2006), for instance, can constrain analyses to previously-validated search spaces and may be most appropriate if the goal is to measure a reliable response across people. However, in some cases, spatial variability in activation might be a key outcome. Validating new tasks by comparing responses to a previously-validated measure in a population you have more control over – for instance, adults in Chapter 2 – may also increase confidence in the novel measure. In the audiobook intervention study (Chapter 6), it was a good sanity check to see that scores on the proximal measures correlated with the standard measures.

It is also important to recognize when the lack of experimental control becomes a barrier to interpreting results. For instance, another fMRI task in the study described in Chapter 4 involved children watching their favorite video clips related to their interest. The control condition was a set of neutral videos depicting scenes in nature. We found huge differences in activation between the personally-interesting videos and the neutral videos, but these differences were much harder to interpret, particularly at the group level. This is because the individual variation in children's personalized videos was enormous: some children's videos involved people, some included music, some

had lots of movement, some were animated… the list goes on. Thus, unlike the language task in which we could impose some experimental control (e.g., always having the same speaker, aiming to match linguistic features as much as possible), the video task was completely uncontrolled, and thus it is much harder to extract a meaningful difference in response. Even within the 'neutral' condition, we found unexpected variation in how much children liked certain videos. After the scan, we asked children how excited they were when each video started in the scanner, for each of their personalized videos and for each of the neutral videos. One of the neutral videos depicting snow falling in a forest tended to have higher ratings than the other neutral videos – sometimes on par with a child's personalized videos. This anecdote has two main lessons: first, that embracing personalization can make analysis *too* complicated, especially with limited data, and second, that it is easy to miss substantial differences in interest and engagement if you do not explicitly look for it.

When does it matter?

Given the resources required, and the slippery slope of theoretical interpretation, it is likely only worth tailoring stimuli in the ways I have described when there is a meaningful payoff. Adults, for instance, will sit through two hours of a monotonous task and will themselves awake for the sake of science – therefore, it may be unnecessary to introduce additional confounds into the stimuli just to make the experience more enjoyable. Toddlers, on the other hand, will not. And that is why this 'ethos of engagement' matters *specifically* for certain populations, including young children.

There are two particular cases that come up in developmental studies in which tailoring stimuli may be especially important: (1) when stimuli design may impact who is studied,

and (2) when stimuli design might impact inferences about individual or group differences.

When stimuli design may impact who is studied

Certain populations are more challenging to study using functional neuroimaging methods. Toddlers are much harder to scan than adults or even older children (Chapter 3). Even when we do collect data from them, the toddlers who comply with our instructions and go in the scanner may not be representative of most toddlers in their age group, limiting the inferences we can draw about typical language development. It is also often harder to scan autistic children than their neurotypical peers, such as when they get frustrated that their favorite videos switch to something boring in a block-design fMRI task and refuse to continue the scan (Chapter 4). And yet, it is even harder to scan minimally verbal autistic children, who are almost never included in fMRI research at all (Tager-Flusberg & Kasari, 2013).

Stepping back for a moment, it is also important to acknowledge that developmental neuroimaging research is also biased toward certain identities, including white, majority-language-speaking, and higher-socioeconomic families (Garcini et al., 2022; Nketia et al., 2021). Of course, this issue of inclusivity is not isolated to developmental neuroimaging (e.g., (Bornstein et al., 2013; Ricard et al., 2023)), nor can it be solved by modifying stimuli. While we made substantial efforts to increase representation in our audiobook intervention – such as by translating all caregiver-facing materials into a second language and hiring bilingual research staff to accommodate Spanish-speaking families – we still could only accommodate two primary languages spoken by parents (English and Spanish), and thus excluded children from language minority homes. Even when we tried to remove barriers to participation through remote administration, the

'digital divide' lingered, and prevented those without necessary technology from being included in research (Chandra et al., 2020; Van Dijk, 2020).

It is tempting, for both scientific and practical reasons, to limit who we study. It is incredibly challenging to disentangle the effects of an audiobook intervention on vocabulary, when children's experiences of schooling, stress, reading difficulty, language background, intervention adherence, and myriad other factors certainly played a role in each individual's outcomes (Chapters 5-6). It was harder to administer an intervention with fidelity when children join on a spotty Wi-Fi network from a cell phone than when a child has their own laptop set up in a home with stable Wi-Fi. It was expensive, time consuming, and messier than if we had limited recruitment to well-off families from a similar background, who were explicitly on the lookout for a program like ours and *able* to comply with the 'ideal' intervention conditions even during a global pandemic. Truly moving towards more representative developmental research requires concerted efforts on many fronts, at all stage of the research process (Bonevski et al., 2014; Bornstein et al., 2013; Garcini et al., 2022; Green et al., 2022; Nketia et al., 2021; Ricard et al., 2023; Webb et al., 2022).

While tailoring stimuli is clearly not the catchall solution, I do think that it is a worthwhile approach to consider in conjunction with other efforts to improve neuroscience through better representation. Across developmental neuroimaging studies in particular, we 'expect' to exclude a substantial proportion of participants due to data quality, compliance, and other factors (Rajagopal et al., 2014; Raschle et al., 2012). If tailoring stimuli to the target population can increase retention, particularly in difficult-to-scan populations, then it may be worth considering.

When stimuli design might impact inferences about individual or group differences

A second situation in which developmental researchers may want to tailor stimuli in some of the ways I previously described is when they are aiming to study individual or group differences.

Neuroscientists have an aspiration to measure brain function and development relate to individual differences. Some researchers have advocated for the use of naturalistic stimuli to study individual differences in brain function (e.g., Finn et al., 2020; Sonkusare et al., 2019; Vanderwal et al., 2019), but the mechanistic explanations for these associations may be difficult to interpret. Linking individual differences in the brain to behavior can be hard (Dubois & Adolphs, 2016), with some calling for much larger sample sizes than are typical in the literature (Marek et al., 2022). Others advocate for maximizing the effect sizes one chooses to study, by maximizing signal and minimizing noise (Gratton et al., 2022). In Chapter 4, we discovered that interest in the material can substantially affect the brain's response to language in children. It is plausible, then, that interest may create noise in an otherwise reliable response to language – or, perhaps, other cognitive functions as well. Particularly in developmental samples, controlling for interest and engagement may be critical for measuring individual differences. Along these same lines, tailoring stimuli may also be important for measuring reliable group differences if heterogeneity in other factors – like interest – might also differ at a group level. In Chapter 4, for instance, we hypothesized that the effect of interest on language processing may have a disproportionate effect on autistic children due to the high prevalence of intense, specific interests (Klin et al., 2007), though we did not find evidence of this group difference in our study.

Future directions

The field of developmental neuroimaging is poised to answer fundamental questions about language development, plasticity, and the role of experience on brain organization. To wrap up this dissertation, here are three questions[30] raised by my work that I think we can address in the future.

1. How does the context of language input impact brain activation during language comprehension, particularly in infants and toddlers?
2. In toddlers, is brain activation during language comprehension driven more by biological maturation (i.e., age) or language skills?
3. By what mechanism does personal interest impact activation in language network in children? Is this only the case for particularly strong interests, or do subtle differences in interest impact neural processing?

These questions highlight the promise of these methods to isolate meaningful individual differences in the brain that relate to behavior, to understand understudied developmental populations, and to explore intersections between language as a cognitive process and the content and context in which it is used during development.

Conclusion

Language development is not only profoundly important to the human experience, but also one of the most striking examples of the interplay between nature and nurture. Studying the neural correlates of this remarkable trajectory is challenging, but it is worth it – we are just beginning to pull back the curtain to reveal the mirror behind it.

[30] Don't worry, I have plenty more.

References

Apthorp, H., Randel, B., Cherasaro, T., Clark, T., McKeown, M., & Beck, I. (2012). Effects of a Supplemental Vocabulary Program on Word Knowledge and Passage Comprehension. *Journal of Research on Educational Effectiveness*, *5*(2), 160–188. https://doi.org/10.1080/19345747.2012.660240

Baldwin, R. S., Peleg-Bruckner, Z., & McClintock, A. H. (1985). Effects of Topic Interest and Prior Knowledge on Reading Comprehension. *Reading Research Quarterly*, *20*(4), 497–504. https://doi.org/10.2307/747856

Bašnáková, J., Weber, K., Petersson, K. M., van Berkum, J., & Hagoort, P. (2014). Beyond the Language Given: The Neural Correlates of Inferring Speaker Meaning. *Cerebral Cortex*, *24*(10), 2572–2578. https://doi.org/10.1093/cercor/bht112

Bedny, M., Pascual-Leone, A., Dodell-Feder, D., Fedorenko, E., & Saxe, R. (2011). Language processing in the occipital cortex of congenitally blind adults. *Proceedings of the National Academy of Sciences*, *108*(11), 4429–4434. https://doi.org/10.1073/pnas.1014818108

Biber, D. (1991). *Variation Across Speech and Writing*. Cambridge University Press.

Bonevski, B., Randell, M., Paul, C., Chapman, K., Twyman, L., Bryant, J., Brozek, I., & Hughes, C. (2014). Reaching the hard-to-reach: A systematic review of strategies for improving health and medical research with socially disadvantaged groups. *BMC Medical Research Methodology*, *14*(1), 42. https://doi.org/10.1186/1471-2288-14-42

Bornstein, M. H., Jager, J., & Putnick, D. L. (2013). Sampling in Developmental Science: Situations, Shortcomings, Solutions, and Standards. *Developmental Review : DR*, *33*(4), 357–370. https://doi.org/10.1016/j.dr.2013.08.003

Cantlon, J. F. (2020). The balance of rigor and reality in developmental neuroscience. *NeuroImage*, *216*, 116464. https://doi.org/10.1016/j.neuroimage.2019.116464

Cantlon, J. F., & Li, R. (2013). Neural Activity during Natural Viewing of Sesame Street Statistically Predicts Test Scores in Early Childhood. *PLOS Biology*, *11*(1), e1001462. https://doi.org/10.1371/journal.pbio.1001462

Chandra, S., Chang, A., Day, L., Fazlullah, A., Liu, J., McBride, L., Mudalige, T., & Weiss, D. (2020). Closing the K–12 digital divide in the age of distance learning. *Common Sense and Boston Consulting Group: Boston, MA, USA*.

Dubois, J., & Adolphs, R. (2016). Building a Science of Individual Differences from fMRI. *Trends in Cognitive Sciences*, *20*(6), 425–443. https://doi.org/10.1016/j.tics.2016.03.014

Elleman, A. M., Lindo, E. J., Morphy, P., & Compton, D. L. (2009). The Impact of Vocabulary Instruction on Passage-Level Comprehension of School-Age Children: A Meta-Analysis. *Journal of Research on Educational Effectiveness*, *2*(1), 1–44. https://doi.org/10.1080/19345740802539200

Emerson, R. W., & Cantlon, J. F. (2012). Early math achievement and functional connectivity in the fronto-parietal network. *Developmental Cognitive Neuroscience*, *2*, S139–S151. https://doi.org/10.1016/j.dcn.2011.11.003

Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New Method for fMRI Investigations of Language: Defining ROIs Functionally in Individual Subjects. *Journal of Neurophysiology*, *104*(2), 1177–1194. https://doi.org/10.1152/jn.00032.2010

Feng, W., Wu, Y., Jan, C., Yu, H., Jiang, X., & Zhou, X. (2017). Effects of contextual relevance on pragmatic inference during conversation: An fMRI study. *Brain and Language*, *171*, 52–61. https://doi.org/10.1016/j.bandl.2017.04.005

Ferjan Ramírez, N., Lytle, S. R., & Kuhl, P. K. (2020). Parent coaching increases conversational turns and advances infant language development. *Proceedings of the National Academy of Sciences*, *117*(7), 3484–3491. https://doi.org/10.1073/pnas.1921653117

Finn, E. S., Glerean, E., Khojandi, A. Y., Nielson, D., Molfese, P. J., Handwerker, D. A., & Bandettini, P. A. (2020). Idiosynchrony: From shared responses to individual differences during naturalistic neuroimaging. *NeuroImage*, *215*, 116828. https://doi.org/10.1016/j.neuroimage.2020.116828

Frew, S., Samara, A., Shearer, H., Eilbott, J., & Vanderwal, T. (2022). Getting the nod: Pediatric head motion in a transdiagnostic sample during movie- and resting-state fMRI. *PLOS ONE*, *17*(4), e0265112. https://doi.org/10.1371/journal.pone.0265112

Garcini, L. M., Arredondo, M. M., Berry, O., Church, J. A., Fryberg, S., Thomason, M. E., & McLaughlin, K. A. (2022). Increasing diversity in developmental cognitive neuroscience: A roadmap for increasing representation in pediatric neuroimaging research. *Developmental Cognitive Neuroscience*, *58*, 101167. https://doi.org/10.1016/j.dcn.2022.101167

Gogate, L. J., & Bahrick, L. E. (1998). Intersensory Redundancy Facilitates Learning of Arbitrary Relations between Vowel Sounds and Objects in Seven-Month-Old Infants. *Journal of Experimental Child Psychology*, *69*(2), 133–149. https://doi.org/10.1006/jecp.1998.2438

Golinkoff, R. M., Can, D. D., Soderstrom, M., & Hirsh-Pasek, K. (2015). (Baby)Talk to Me: The Social Context of Infant-Directed Speech and Its Effects on Early

Language Acquisition. *Current Directions in Psychological Science*, *24*(5), 339–344. https://doi.org/10.1177/0963721415595345

Gratton, C., Nelson, S. M., & Gordon, E. M. (2022). Brain-behavior correlations: Two paths toward reliability. *Neuron*, *110*(9), 1446–1449. https://doi.org/10.1016/j.neuron.2022.04.018

Green, K. H., Van De Groep, I. H., Te Brinke, L. W., van der Cruijsen, R., van Rossenberg, F., & El Marroun, H. (2022). A perspective on enhancing representative samples in developmental human neuroscience: Connecting science to society. *Frontiers in Integrative Neuroscience*, *16*. https://www.frontiersin.org/articles/10.3389/fnint.2022.981657

Hirsh-Pasek, K., Adamson, L. B., Bakeman, R., Owen, M. T., Golinkoff, R. M., Pace, A., Yust, P. K. S., & Suma, K. (2015). The Contribution of Early Communication Quality to Low-Income Children's Language Success. *Psychological Science*, *26*(7), 1071–1083. https://doi.org/10.1177/0956797615581493

Hoff, E. (2006). How social contexts support and shape language development. *Developmental Review*, *26*(1), 55–88. https://doi.org/10.1016/j.dr.2005.11.002

Honey, C. J., Thompson, C. R., Lerner, Y., & Hasson, U. (2012). Not Lost in Translation: Neural Responses Shared Across Languages. *Journal of Neuroscience*, *32*(44), 15277–15283. https://doi.org/10.1523/JNEUROSCI.1800-12.2012

Jang, G., Yoon, S., Lee, S.-E., Park, H., Kim, J., Ko, J. H., & Park, H.-J. (2013). Everyday conversation requires cognitive inference: Neural bases of comprehending implicated meanings in conversations. *NeuroImage*, *81*, 61–72. https://doi.org/10.1016/j.neuroimage.2013.05.027

Kamps, F. S., Richardson, H., Murty, N. A. R., Kanwisher, N., & Saxe, R. (2022). Using child-friendly movie stimuli to study the development of face, place, and object regions from age 3 to 12 years. *Human Brain Mapping*, *43*(9), 2782–2800. https://doi.org/10.1002/hbm.25815

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *Journal of Neuroscience*, *17*(11), 4302–4311. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997

Kendeou, P., Rapp, D. N., & van den Broek, P. (2003). The influence of reader's prior knowledge on text comprehension and learning from text. In R. Nata (Ed.), *Progress in Education* (Vol. 13, pp. 189–209). Nova Science Publishers, Inc.

Klin, A., Danovitch, J. H., Merz, A. B., & Volkmar, F. R. (2007). Circumscribed Interests in Higher Functioning Individuals With Autism Spectrum Disorders: An Exploratory Study. *Research and Practice for Persons with Severe Disabilities*, *32*(2), 89–100. https://doi.org/10.2511/rpsd.32.2.89

Kominsky, J. F., Lucca, K., Thomas, A. J., Frank, M. C., & Hamlin, J. K. (2022). Simplicity and validity in infant research. *Cognitive Development*, *63*, 101213. https://doi.org/10.1016/j.cogdev.2022.101213

Kosakowski, H. L., Cohen, M. A., Takahashi, A., Keil, B., Kanwisher, N., & Saxe, R. (2022). Selective responses to faces, scenes, and bodies in the ventral visual pathway of infants. *Current Biology*, *32*(2), 265-274.e5. https://doi.org/10.1016/j.cub.2021.10.064

Kuhl, P. K. (2007). Is speech learning 'gated' by the social brain? *Developmental Science*, *10*(1), 110–120. https://doi.org/10.1111/j.1467-7687.2007.00572.x

Lewkowicz, D. J., & Flom, R. (2014). The Audiovisual Temporal Binding Window Narrows in Early Childhood. *Child Development*, *85*(2), 685–694. https://doi.org/10.1111/cdev.12142

MacSweeney, M., Capek, C. M., Campbell, R., & Woll, B. (2008). The signing brain: The neurobiology of sign language. *Trends in Cognitive Sciences*, *12*(11), 432–440. https://doi.org/10.1016/j.tics.2008.07.010

Malik-Moraleda, S., Ayyash, D., Gallée, J., Affourtit, J., Hoffmann, M., Mineroff, Z., Jouravlev, O., & Fedorenko, E. (2022). An investigation across 45 languages and 12 language families reveals a universal language network. *Nature Neuroscience*, *25*(8), Article 8. https://doi.org/10.1038/s41593-022-01114-5

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., … Dosenbach, N. U. F. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, *603*(7902), Article 7902. https://doi.org/10.1038/s41586-022-04492-9

Mody, M., & Belliveau, J. W. (2013). Speech and Language Impairments in Autism: Insights from Behavior and Neuroimaging. *North American Journal of Medicine & Science*, *5*(3), 157–161.

Montag, J. L., Jones, M. N., & Smith, L. B. (2015). The Words Children Hear: Picture Books and the Statistics for Language Learning. *Psychological Science*, *26*(9), 1489–1496. https://doi.org/10.1177/0956797615594361

Montag, J. L., & MacDonald, M. C. (2015). Text exposure predicts spoken production of complex sentences in 8- and 12-year-old children and adults. *Journal of Experimental Psychology: General*, *144*, 447–468. https://doi.org/10.1037/xge0000054

Neville, H. J., Bavelier, D., Corina, D., Rauschecker, J., Karni, A., Lalwani, A., Braun, A., Clark, V., Jezzard, P., & Turner, R. (1998). Cerebral organization for language in deaf and hearing subjects: Biological constraints and effects of experience.

Proceedings of the National Academy of Sciences, *95*(3), 922–929. https://doi.org/10.1073/pnas.95.3.922

Nketia, J., Amso, D., & Brito, N. H. (2021). Towards a more inclusive and equitable developmental cognitive neuroscience. *Developmental Cognitive Neuroscience*, *52*, 101014. https://doi.org/10.1016/j.dcn.2021.101014

Obretenova, S., Halko, M., Plow, E., Pascual-Leone, A., & Merabet, L. (2010). Neuroplasticity associated with tactile language communication in a deaf-blind subject. *Frontiers in Human Neuroscience*, *3*. https://www.frontiersin.org/articles/10.3389/neuro.09.060.2009

Packer, M. J., & Moreno-Dulcey, F. A. (2022). Theory of puppets?: A critique of the use of puppets as stimulus materials in psychological research with young children. *Cognitive Development*, *61*, 101146. https://doi.org/10.1016/j.cogdev.2021.101146

Paunov, A. M., Blank, I. A., & Fedorenko, E. (2019). Functionally distinct language and Theory of Mind networks are synchronized at rest and during language comprehension. *Journal of Neurophysiology*, *121*(4), 1244–1265. https://doi.org/10.1152/jn.00619.2018

Paunov, A. M., Blank, I. A., Jouravlev, O., Mineroff, Z., Gallée, J., & Fedorenko, E. (2022). Differential Tracking of Linguistic vs. Mental State Content in Naturalistic Stimuli by Language and Theory of Mind (ToM) Brain Networks. *Neurobiology of Language*, 1–29. https://doi.org/10.1162/nol_a_00071

Rajagopal, A., Byars, A., Schapiro, M., Lee, G. R., & Holland, S. K. (2014). Success Rates for Functional MR Imaging in Children. *American Journal of Neuroradiology*, *35*(12), 2319–2325. https://doi.org/10.3174/ajnr.A4062

Raschle, N., Zuk, J., Ortiz-Mantilla, S., Sliva, D. D., Franceschi, A., Grant, P. E., Benasich, A. A., & Gaab, N. (2012). Pediatric neuroimaging in early childhood and infancy: Challenges and practical guidelines. *Annals of the New York Academy of Sciences*, *1252*(1), 43–50. https://doi.org/10.1111/j.1749-6632.2012.06457.x

Recht, D. R., & Leslie, L. (1988). Effect of prior knowledge on good and poor readers' memory of text. *Journal of Educational Psychology*, *80*, 16–20. https://doi.org/10.1037/0022-0663.80.1.16

Redcay, E., & Moraczewski, D. (2020). Social cognition in context: A naturalistic imaging approach. *NeuroImage*, *216*, 116392. https://doi.org/10.1016/j.neuroimage.2019.116392

Ricard, J. A., Parker, T. C., Dhamala, E., Kwasa, J., Allsop, A., & Holmes, A. J. (2023). Confronting racially exclusionary practices in the acquisition and analyses of

neuroimaging data. *Nature Neuroscience*, *26*(1), Article 1. https://doi.org/10.1038/s41593-022-01218-y

Richardson, H., Koster-Hale, J., Caselli, N., Magid, R., Benedict, R., Olson, H., Pyers, J., & Saxe, R. (2020). Reduced neural selectivity for mental states in deaf children with delayed exposure to sign language. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-17004-y

Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., & Saxe, R. (2018). Development of the social brain from age three to twelve years. *Nature Communications*, *9*(1), Article 1. https://doi.org/10.1038/s41467-018-03399-2

Romeo, R. R., Leonard, J. A., Robinson, S. T., West, M. R., Mackey, A. P., Rowe, M. L., & Gabrieli, J. D. E. (2018). Beyond the 30-Million-Word Gap: Children's Conversational Exposure Is Associated With Language-Related Brain Function. *Psychological Science*, *29*(5), 700–710. https://doi.org/10.1177/0956797617742725

Rowe, M. L. (2008). Child-directed speech: Relation to socioeconomic status, knowledge of child development and child vocabulary skill*. *Journal of Child Language*, *35*(1), 185–205. https://doi.org/10.1017/S0305000907008343

Saxe, R., Brett, M., & Kanwisher, N. (2006). Divide and conquer: A defense of functional localizers. *NeuroImage*, *30*(4), 1088–1096. https://doi.org/10.1016/j.neuroimage.2005.12.062

Scott, T. L., Gallée, J., & Fedorenko, E. (2017). A new fun and robust version of an fMRI localizer for the frontotemporal language system. *Cognitive Neuroscience*, *8*(3), 167–176. https://doi.org/10.1080/17588928.2016.1201466

Seyfarth, R. M., & Cheney, D. L. (2014). The evolution of language from social cognition. *Current Opinion in Neurobiology*, *28*, 5–9. https://doi.org/10.1016/j.conb.2014.04.003

Shnayer, S. W. (1968). *Some Relationships between Reading Interest and Reading Comprehension*. https://eric.ed.gov/?id=ED022633

Sonkusare, S., Breakspear, M., & Guo, C. (2019). Naturalistic Stimuli in Neuroscience: Critically Acclaimed. *Trends in Cognitive Sciences*, *23*(8), 699–714. https://doi.org/10.1016/j.tics.2019.05.004

Tager-Flusberg, H., & Kasari, C. (2013). Minimally Verbal School-Aged Children with Autism Spectrum Disorder: The Neglected End of the Spectrum. *Autism Research*, *6*(6), 468–478. https://doi.org/10.1002/aur.1329

Vadasy, P. F., Sanders, E. A., & Logan Herrera, B. (2015). Efficacy of Rich Vocabulary Instruction in Fourth- and Fifth-Grade Classrooms. *Journal of Research on Educational Effectiveness*, *8*(3), 325–365. https://doi.org/10.1080/19345747.2014.933495

Van Dijk, J. (2020). *The digital divide*. John Wiley & Sons.

Vanderwal, T., Eilbott, J., & Castellanos, F. X. (2019). Movies in the magnet: Naturalistic paradigms in developmental functional neuroimaging. *Developmental Cognitive Neuroscience*, *36*, 100600. https://doi.org/10.1016/j.dcn.2018.10.004

Webb, E. K., Cardenas-Iniguez, C., & Douglas, R. (2022). Radically reframing studies on neurobiology and socioeconomic circumstances: A call for social justice-oriented neuroscience. *Frontiers in Integrative Neuroscience*, *16*, 958545. https://doi.org/10.3389/fnint.2022.958545

Yates, T. S., Skalaban, L. J., Ellis, C. T., Bracher, A. J., Baldassano, C., & Turk-Browne, N. B. (2022). Neural event segmentation of continuous experience in human infants. *Proceedings of the National Academy of Sciences*, *119*(43), e2200257119. https://doi.org/10.1073/pnas.2200257119