Children seek visual information during signed and spoken language comprehension

Kyle MacDonald[1], Virginia Marchman[1], Anne Fernald[1], & Michael C. Frank[1]

[1] Stanford University

Author Note

Correspondence concerning this article should be addressed to Kyle MacDonald, 450 Serra Mall, Stanford, CA 94306. E-mail: kylem4@stanford.edu

Abstract

Understanding grounded language involves linking the incoming linguistic signal to the visual world. Information that is gathered through visual fixations can facilitate the comprehension process. But how do listeners decide what visual information to gather? Here, we propose that children flexibly adapt their gaze to seek visual information from social partners to support language understanding. We present evidence for our explanation using two case studies of eye movements during real-time language processing. First, compared to children learning spoken English (n=80), young ASL-learners (n=30) delayed gaze shifts away from a language source, were more accurate and produced a higher proportion of language-driven shifts. Second, English-speaking preschoolers (n=39) and adults (n=31) delayed the timing of gaze shifts away from a speaker's face while processing language in a noisy environment. This delay resulted in more accurate responses and a higher proportion of language-driven gaze shifts. These results suggest that young listeners can adapt their gaze to seek supportive visual information from their social partners during real-time language comprehension.

Keywords: eye movements; grounded language comprehension; information-seeking; speech in background noise; American Sign Language

Word count: X

Children seek visual information during signed and spoken language comprehension

Introduction

Extracting meaning from language represents a formidable challenge for young learners. Consider that even in the simple case of understanding grounded, familiar language (e.g., "look at the ball"), listeners must integrate linguistic and non-linguistic information from two continuous streams of input. Moreover, language unfolds within dynamic interactions where there is often insufficient information to figure out what is being said, and yet listeners must decide how best to respond. Despite these challenges, even young children are capable of linking language to the world quite efficiently, shifting visual attention to a named object in a scene within hundreds of milliseconds upon hearing it's name (Allopenna, Magnuson, & Tanenhaus, 1998; Spivey, Tanenhaus, Eberhard, & Sedivy, 2002; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). How do children comprehend language despite noisy input and limited processing capabilities?

One solution is for the listener to integrate multiple sources of information to constrain the set of possible interpretations of an utterance (MacDonald & Seidenberg, 2006; McClelland & Elman, 1986). Under this more interactive account, listeners comprehend words by partially activating several candidates that are consistent with incoming perceptual information. Then, as more information arrives, words that do not match the perceptual signal are no longer considered, and words that do match become more strongly activated until a single interpretation is reached (see McClelland, Mirman, and Holt (2006) for a review). Critically, information from multiple sources – the linguistic signal, visual world, and conceptual knowledge – mutually influences one another to constrain the listener's interpretation of an utterance. For example, if a speaker's mouth movements suggest one sound while their acoustic output indicates another, the interaction results in the listener perceiving a third, intermediate sound (the "McGurk effect") (MacDonald & McGurk, 1978).

Other research in psycholinguistics shows that adults can use information in the visual scene to help them parse syntactically ambiguous utterances (Tanenhaus et al., 1995).

Thus, information gathered from the visual world can facilitate language comprehension. The incoming linguistic signal, however, is ephemeral and multiple fixation behaviors could be useful, meaning that listeners must quickly decide how to direct their gaze to informative locations in the environment. Consider a speaker who asks you to "Pass the salt" in a noisy restaurant where it is difficult to perceive what she is saying. Recent theoretical and empirical work suggests that children and adults handle this sort of noise in the signal by integrating what they percieve with their prior beliefs about the plausibility of a speaker's intended meaning (Fourtassi & Frank, 2017; Gibson, Bergen, & Piantadosi, 2013; Yurovsky, Case, & Frank, 2017). Here, we pursue the idea that comprehension could also be supported by integrating visual information gathered via looks to different locations in the visual scene. Returning the noisy restaurant example, the listener could support comprehension by looking at the objects on the table (e.g., the type of food the speaker is eating), or by looking at the speaker directly (e.g., reading her lips or the direction of her gaze).

Another interesting case where there is competition for visual attention is the comprehension of a visual-manual language such as American Sign Language (ASL). In ASL, fixations to a signer are highly informative because all linguistic information is processed via the visual channel. Moreover, the decision to look away from a signer to the rest of the visual world could be risky because this behavior might reduce visual access to subsquent linguistic information, thus complicating the listener's decision to gather visual information that might constrain the meaning of a noisy utterance. Finally, ASL represents an interesting case where there is direct competition between allocating visual attention to the non-linguistic information – gesture, eye gaze, and facial expressions – that has been suggested to support spoken language comprehension.

Taken together, the noisy restaurant and ASL examples highlight how eye movements can be chracterized as an active decision-making process where listeners select fixations to gather language-relevant information. In the current work, we pursue this idea and propose that listeners are sensitive to the value of different fixation behaviors with respect to the goal of grounded language understanding. We hypothesize that even young children can flexibly adapt the dynamics of their gaze to seek higher value visual information that supports comprehension. This hypothesis is inspired by ideas from several research programs, including work on language-driven shifts in visual attention (Allopenna et al., 1998; Tanenhaus et al., 1995), goal-based accounts of eye movements in everyday tasks (Hayhoe & Ballard, 2005), and language perception as a process of multisensory cue integration (Vigliocco, Perniss, & Vinson, 2014). In the following sections, we briefly review each of these literatures to motivate our explanation of information-seeking eye movements in grounded signed and spoken language comprehension.

Vision-language interactions during language comprehension

The study of eye movements during spoken language comprehension has provided insight into the interaction between concepts, language, and visual attention. The majority of this work has used the Visual World Paradigm (VWP) where listeners' eye movements are recorded at the millisecond timescale while processing language and looking at a set of objects (see Salverda, Brown, and Tanenhaus (2011) for a review). Crucially, these analyses rely on the fact that listeners will initiate gaze shifts to named referents with only partial information, in contrast to waiting until the end of a cognitive process (Gold & Shadlen, 2000). Thus, the timecourse of eye movements can provide a window onto how and when people integrate information to reach an interpretation of the incoming linguistic signal.

A classic finding using the VWP shows that listeners will rapidly shift visual attention upon hearing the name of an object ("Pick up a beaker.") in the visual scene with a high proportion of shifts occurring soon after the target word begins (Allopenna et al., 1998). Moreover, adults will look at phonological onset-competitor ("beetle") early upon hearing the word "beaker," suggesting that they activate multiple interpretations and resolve ambiguity as the stimulus unfolds. These results fall out of predictions made by a family of interactive models of speech perception where information from multiple sources is integrated rapidly to constrain language understanding (MacDonald & Seidenberg, 2006; McClelland et al., 2006).

The visual world can also constrain the set of plausible interpretations of language (Dahan & Tanenhaus, 2005; Yee & Sedivy, 2006). For example, Altmann and Kamide (2007) showed that people will allocate more looks to an empty wine glass as compared to a full beer glass upon hearing the past tense verb "has drunk." They proposed that anticipatory eye movements reflect the ability of the visual information in a scene to activate listeners' conceptual representations prior to the arrival of the linguistic signal (see also Huettig and Altmann, 2005).

In addition to work in adult psycholinguistics, the VWP has been useful for studying developmental change in language comprehension skill in children. Researchers have adapted the task to measure the timing and accuracy of children's gaze shifts as they look at two familiar objects and listen to simple sentences naming one of the objects (Fernald, Zangl, Portillo, & Marchman, 2008; Venker, Eernisse, Saffran, & Weismer, 2013). Such research finds that children, like adults, shift gaze to named objects occur soon after the auditory information is sufficient to enable referent identification. Moreover, individual differences in the speed and accuracy of eye movements predict vocabulary growth and later language and cognitive outcomes (Fernald, Perfors, & Marchman, 2006; Marchman & Fernald, 2008; Rigler et al., 2015).

Goal-based accounts of eye movements in everyday tasks

The majority of the work on language-driven visual attention has used eye movements as an index of the underlying interaction between linguistic, visual, and conceptual information. This approach reflects a somewhat passive construal of how listeners might allocate visual attention during language comprehension. A parallel body of work on goal-based accounts of vision starts from the idea that eye movements reflect an active information-gathering process during which visual fixations are driven by task goals (Hayhoe & Ballard, 2005).

Under these accounts, perceivers deploy their gaze to reduce uncertainty about the world and to maximize their expected future rewards with respect to some goal. For example, Hayhoe and Ballard (2005) review evidence that adults fixate on locations that are most helpful for their current task (e.g., looks to an upcoming obstacle when walking) as opposed to other aspects of a visual scene that might be more salient (e.g., a flashing light). Moreover, empirical work shows that adults gather task-specific information via different visual routines as they become useful for their goals. For example, Triesch, Ballard, Hayhoe, and Sullivan (2003) found that adults were less likely to gather and store visual information about the size of an object when it was not relevant to the task of sorting and stacking the objects.

Hayhoe and Ballard (2005)'s review also highlights how perceivers learn to deploy efficient gaze patterns over the course of becoming more familiar with a task. They point out that visual routines are developed over time, and it is only when a task becomes highly-practiced that people allocate fewer looks to less-relevant aspects of the scene. For example, Shinoda, Hayhoe, and Shrivastava (2001) show that skilled drivers learn to spread visual attention more broadly at intersections to better detect stop signs. Other empirical work shows that the visual system rapidly learns to use temporal regularities in the environment to control the timing of eye movements to detect goal-relevant events (Hoppe &

Rothkopf, 2016). Finally, the timing of eye movements in these tasks often occur before an expected event, suggesting that gaze patterns reflect an interaction between people's expectations, the information available in the visual scene, and their task goals.

Connecting this body of evidence with the research on visual language processing reviewed above, recent theoretical work has argued for a stronger link between goal-based perspectives and work on eye movements during language comprehension. Salverda et al. (2011) highlight the immediate relevance of visual information for language understanding, suggesting that listeners' goals should be a key predictor of fixation behaviors. Moreover, they point out that factors such as the difficulty of executing a real world task should change decisions about where to look during comprehension. One example of starting from a goal-based approach comes from Nelson and Cottrell (2007)'s study of gaze patterns during category learning. Nelson and Cottrell (2007) modeled eye movements as a type of question-asking behavior and found that when participants became more familiar with novel concepts, their gaze patterns shifted from exploratory to efficient, suggesting that fixations changed as a function of goals during the task.

Pursuing this connection further, in our current studies, goal-based models of eye movements predict that gaze during language comprehension should adapt to the processing context. That is, listeners should change the timing and location of eye movements when a fixation area become more useful for comprehension. This proposal, which we test, dovetails with a growing body of research that explores the effects of multisensory (gesture, prosody, facial expression and body movement) integration on language perception and comprehension.

Language perception as multisensory integration

The final line of research that informs our studies is work exploring the process of language comprehension as multisenory integration. This research starts from the idea that language understanding does not just involve a single stream of linguistic information. Instead, face-to-face communication provides the listener with access to a set of multimodal cues that can shape language understanding (for a review, see Vigliocco, Perniss, and Vinson, 2014). For example, empirical work shows that when gesture and speech provide redundant cues to meaning, adults are faster to process the information and make fewer comprehension errors (Kelly, Özyürek, & Maris, 2010). Moreover, developmental work shows that parents use visual cues such as gesture and eye gaze to structure language interactions with their children (Estigarribia & Clark, 2007). And, from a young age, children also produce gestures such as reaches and points to share attention with others to achieve communicative goals (Liszkowski, Brown, Callaghan, Takada, & De Vos, 2012).

In fact, most developmental accounts of early language acquisition begin from the ecological context of children grounding language within multimodal, social interations (Clark, 2009; Tomasello & Farrar, 1986). This literature has often focused on how children integrate social cues that are processed in a modality different from the linguistic signal (i.e., spoken words are processed via audition while a speaker's eye gaze or points are processed via vision). The case of ASL, which we explore in this work, highlights how the process of integrating social cues with the linguistic signal is not necessarily cross-modal. When children comprehend ASL, both signs and social cues are visual and could compete for fixations. And yet we know little about how young ASL-learners deploy fixations to gather information about signs, social cues, or the contents of the visual world.

Additional support for the role of multisensory processing in language comes from work on audiovisual speech perception. These studies show that spoken language perception can be shaped by visual information coming from a speaker's mouth. In a review, Peelle and Sommers (2015) point out that mouth movements provide a clear indication of when someone has started to speak, which cues the listener to allocate additional attention to the speech signal. Moreover, a speaker's mouth movements convey information about the phonemes in the acoustic signal. For example, visual speech information distinguishes between consonants such as **/b/** vs. **/d/** and place of articulation can help a listener differentiate between words such as "cat" or "cap." Finally, classic empirical work shows benefits for audiovisual speech perception compared to auditory- or visual-only speech perception, especially in noisy listening contexts (Erber, 1969).

In sum, work on multisensory processing shows that auditory and visual information interact to shape language perception. These results parallel the predictions of interactive models of language processing reviewed earlier (MacDonald & Seidenberg, 2006; McClelland et al., 2006), and they suggest that visual information from a social partner should be considered as an input to children's language comprehension. Finally, this work highlights the importance of studying comprehension within face-to-face communication, where listeners can choose to look at their social partners to gather language-releveant information.

The present studies

In the studies reported here, we explore an information-seeking explanation of eye movements during grounded signed and spoken language comprehension. We propose that the timing of gaze shifts is related to the goal of gathering language-relevant visual information from a speaker balanced with fixating on the surrounding visual scene. We draw on models of eye movements as active decisions that gather information to achieve reliable interpretations of incoming language and test predictions of our account using two case studies: processing of signed vs. spoken language and processing spoken language in noisy vs. clear auditory environments. These cases, while superficially different, share a key feature: The interaction between the listener and their environment changes the value of fixating on the source of language to support comprehension. For example, in comparing ASL to spoken language, the value of looking to an interloctuor is higher since all of the language-relevant information is located at that point in the visual world; whereas a young child processing spoken language can fixate on other locations in the visual scene while still processing linguistic information via the auditory channel.

A secondary goal of this work was to test whether children and adults would show similar patterns of gaze adaptation in response to changes in the value of looking to a social partner for language understanding. Recent developmental work shows that, like adults, preschoolers will flexibly adjust how they interpret ambiguous sentences (e.g., "I had carrots and *bees* for dinner.") by integrating information about the reliability of the incoming perceptual information with their expectations about the speaker (Gibson et al., 2013; Yurovsky et al., 2017). While children's behavior paralleled adults, they relied more on top-down expectations about the speaker, perhaps because their perceptual representations were noisier. These developmental differences provide insight into how children succeed in understanding language despite having partial knowledge of word-object mappings.

The key behavioral prediction is that children and adults will adapt the timing of their eye movements to facilitate word recognition. We hypothesized that as fixations to the source of language – either a signer or a speaker – provide higher value visual information, listeners should prioritize looking to their social partner. Concretely, in a noisy auditory environment, looks to a speaker's face should be more useful compared to the same behavior without background noise where it is easier to perceive the auditory signal. In this case, we predict that listeners would be (a) slower to shift gaze away from the speaker's face, which in turn would lead to (b) more consistent shifts to named objects and (c) fewer early, nonlanguage-driven eye movements to the *rest* of the visual world.

The structure of the paper is as follows. First, we describe and motivate our analytic approach. Then, we present a comparison of the timing and accuracy of children's eye movements while they processed signed vs. spoken language. Finally, we compare both children and adults' gaze patterns as they processed speech in either noisy vs. clear auditory environments.

Analytic approach

Before describing the studies, it is worth motivating our analytic approach. To quantify evidence for our predictions, we use four analyses: (1) the timecourse of listeners' looking to each area of interest (AOI), (2) the reaction time (RT) and accuracy of listeners' first shifts away from the signer/speaker, (3) an Exponentially Weighted Moving Average (EWMA) of first shifts, and (4) a Drift Diffusion Model (DDM) of first shifts.[1]

For each experiment, we first present an analysis of the timecourse of participants' looking to each AOI in the visual scene as the target sentence unfolded. Proportion looking reflects the mean proportion of trials on which participants fixated on the signer/speaker, the target image, or the distracter image at every 33-ms interval of the stimulus sentence. We tested condition differences in the proportion looking to the language source – signer or speaker – using a nonparametric cluster-based permutation analysis, which accounts for the issue of taking multiple comparisons across many time bins in the timecourse (Maris & Oostenveld, 2007). It is important to note that this analysis tests binary hypotheses of differences between time sereies and does not provide evidence about the precise timing of those differences.

Next, we analyzed the RT and accuracy of participants' first gaze shifts away from the signer/speaker. RT corresponds to the latency of shifting gaze away from the central stimulus to either object measured from the onset of the target noun. All reaction time

[1] All analysis code can be found in the online repository for this project: https://github.com/kemacdonald/speed-acc.

distributions were trimmed to between zero and two seconds and RTs were modeled in log space. Accuracy corresponds to whether participants' first gaze shift landed on the target or the distracter object. It is important to note that this analysis of accuracy does not focus on the amount of time spent looking at the target vs. the distracter image – a measure typically used in analyses of the Visual World Paradigm. We chose to analyze first shifts because we think that they reflect rapid decisions driven by accumulating information about the identity of the named object, and thus provide a window onto changes in the underlying dynamics of how listeners integrate linguistic and visual information.

We used the `rstanarm` (Gabry & Goodrich, 2016) package to fit Bayesian mixed-effects regression models. The mixed-effects approach allowed us to model the nested structure of our data – multiple trials for each participant and item, and a within-participants manipulation – by including random intercepts for each participant and item, and a random slope for each item and noise condition. We used Bayesian estimation to quantify uncertainty in our point estimates, which we communicate using a 95% Highest Density Interval (HDI). The HDI provides a range of credible values given the data and model. Finally, to estimate age-related differences, we fit two types of models: (1) age group (adults vs. children) as a categorical predictor and (2) age as a continuous predictor (measured in days) within the child sample. In the main text, we report specific effects and contrasts of interest for our hypotheses, but, in the Appendix, we report the full model output for each analytic model in the paper.

Following the behavioral results, we present two model-based analyses. The goal of each model is to move beyond a description of the data and to map behavioral differences to underlying psychological processes. The EWMA models changes in the tendency to generate random gaze shifts as a function of when they occurred in the RT distribution (Vandekerckhove & Tuerlinckx, 2007). For each RT, the model generates two values: a "control statistic" (CS, which captures the running average accuracy of first shifts) and an "upper control limit" (UCL, which captures the pre-defined limit of when accuracy would be categorized as better than guessing). Here, the CS is an expectation of random shifting to either the target or the distracter image (nonlanguage-driven shifts), or a Bernoulli process with probability of success 0.5. As RTs get slower, we assume that participants have gathered more information and should become more accurate (i.e., language-driven), or a Bernoulli process with probability success > 0.5. Using this model, we can quantify the proportion of gaze shifts that were classified as language-driven as opposed to guessing.

Finally, following Vandekerckhove and Tuerlinckx (2007), we selected the shifts categorized as language-driven by the EWMA and fit a hierarchical Bayesian Drift-Diffusion Model (HDDM). The DDM is a cognitive model of decision making developed over the past forty years (Ratcliff & McKoon, 2008) that can help to quantify differences in the underlying decision process that lead to different patterns of observable behavior. The model assumes that people accumulate noisy evidence in favor of one alternative with a response generated when the evidence crosses a pre-defined decision threshold. We chose to implement a hierarchical Bayesian version of the DDM using the HDDM Python package (Wiecki, Sofer, & Frank, 2013) since we had relatively few trials from child participants and recent simulation studies have shown that the HDDM approach was better than other fitting methods for small data sets (Ratcliff & Childers, 2015). Here, we focus on two parameters of interest for our hypotheses: *boundary separation*, which indexes the amount of evidence gathered before generating a response (higher values suggest more cautious responding) and *drift rate*, which indexes the amount of evidence accumulated per unit time (higher values suggest more efficient processing).

Experiment 1

In Experiment 1, we compared eye movements of children learning American Sign Language to children learning a spoken language using parallel real-time language

Table 1

Age distributions of children in Experiment 1. All ages are reported in months.

Center Stimulus	Mean	Min	Max	n
ASL	27.90	16.00	53.00	30.00
Face	26.00	25.00	26.00	24.00
Object	31.90	26.00	39.00	40.00
Bullseye	26.10	26.00	27.00	16.00

comprehension tasks. In the task, children processed familiar sentences (e.g., "Where's the ball?") while looking at a simplified visual world with three fixation targets (a center stimulus that varied by condition, a target picture, and a distracter picture; see Figure 1). The spoken language data are a reanalysis of three unpublished data sets, and the ASL data are reported in MacDonald, LaMarr, Corina, Marchman, and Fernald (2018). Our primary question of interest is whether processing a sign language like ASL would increase the value of fixating on the language source and decrease the value of generating exploratory, nonlanguage-driven shifts even after the disambiguation point in the linguistic signal. If ASL learners are sensitive to the cost of shifting gaze away from a signer, then they would show evidence of prioritizing accuracy over and above speed of shifting gaze to the named object.

Methods

Participants. Table 1 contains details about the age distributions of children in all four samples.

Spoken English samples. Participants were 80 native, monolingual English-learning children divided across three samples with no reported history of developmental or language

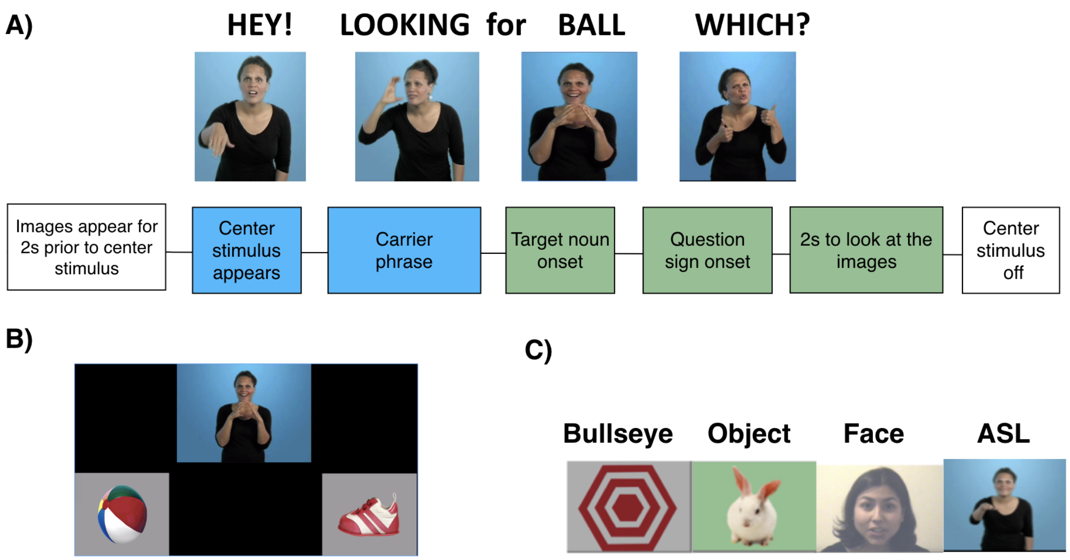

Figure 1. Stimuli for Experiments 1 and 2. Panel A shows the timecourse of the linguistic stimuli for a single trial for children learning American Sign Language. Panel B shows the layout of the fixation locations for all tasks: the center stimulus, the target, and the distracter. Panel C shows the four center stimulus items: a static geometric shape (Bullseye), a static image of a familiar object (Object), a person speaking (Face), a person signing (ASL).

delay.

ASL sample. Participants were 30 native, monolingual ASL-learning children (18 deaf, 12 hearing). All children, regardless of hearing status, were exposed to ASL from birth through extensive interaction with at least one caregiver fluent in ASL and were reported to experience at least 80% ASL in their daily lives. The ASL sample included a wider age range compared to the spoken English samples because this is a rare population.

Stimuli. There are differences between ASL and English question structures. However, all linguistic stimuli shared the same trial structure: language to attract participants' attention followed by a sentence containing a target noun.

ASL linguistic stimuli. We recorded two sets of ASL stimuli, using two valid ASL sentence structures for questions: 1) Sentence-initial wh-phrase: “HEY! WHERE [target noun]?” and 2) Sentence-final wh-phrase: “HEY! [target noun] WHERE?” Two female native ASL users recorded several tokens of each sentence in a child-directed register. Before each sentence, the signer produced a common attention-getting gesture. Mean sign length was 1254 ms, ranging from 693 ms to 1980 ms

English linguistic stimuli. All three tasks (Object, Bullseye, and Face) featured the same female speaker who used natural child-directed speech and said: “Look! Where’s the (target word)?” The target words were: ball, banana, book, cookie, juice, and shoe. For the Face task, a female native English speaker was video-recorded as she looked straight ahead and said, “Look! Where’s the (target word)?” Mean word length was 786.70 ms, ranging from 600 ms to 940 ms.

ASL and English visual stimuli. The image set consisted of colorful digitized pictures of objects presented in fixed pairs with no phonological overlap (ASL task: cat—bird, car—book, bear—doll, ball—shoe; English tasks: book-shoe, juice-banana, cookie-ball). Side of target picture was counterbalanced across trials.

Design and procedure. *Trial structure.* Children sat on their caregiver’s lap and viewed the task on a screen while their gaze was recorded using a digital camcorder. On each trial, the child saw two images of familiar objects on the screen for two seconds before the center stimulus appeared. This time allowed the child to visually explore both images. Next, the target sentence – which consisted of a carrier phrase, target noun, and question sign – was presented, followed by two seconds without language to allow the child to respond to the signer’s sentence. The trial structure of the Face, Object, and Bullseye tasks were highly similar: children were given two seconds to visually explore the objects prior to the appearance of the center stimulus, then processed a target sentence, and finally were given two seconds of silence to generate a response to the target noun. Participants saw 32 test trials with several filler trials interspersed to maintain interest.

Coding. Participants' gaze patterns were videotaped and later coded frame-by-frame at 33-ms resolution by trained coders blind to target side. On each trial, coders indicated whether the eyes were fixated on the central signer, one of the images, shifting between pictures, or away (off), yielding a high-resolution record of eye movements aligned with target noun onset. Prior to coding, all trials were pre-screened to exclude those few trials on which the participant was inattentive or there was external interference. To assess inter-coder reliability, 25% of the videos were re-coded. Agreement was scored at the level of individual frames of video and averaged 98% on these reliability assessments.

Results

Behavioral analyses. *Timecourse looking.* The first question of interest was how do young ASL and English learners distribute attention across the three fixation locations while processing language in real-time? Figure 2A presents an overview of children's looking to each AOI for each processing context. This plot shows changes in the mean proportion of trials on which participants fixated the center stimulus, the target image, or the distracter image at every 33-ms interval of the stimulus sentence. At target-noun onset, children tended to look at the center stimulus. As the target noun unfolded, the mean proportion looking to the center decreased rapidly as participants shifted their gaze to the target or the distracter image. Proportion looking to the target increased sooner and reached a higher asymptote compared to proportion looking to the distracter for all four contexts.

After looking to the target image, participants tended to shift their gaze back to the center, shown by the increase in proportion looking to the center around two seconds after target-noun onset. There were several qualitative differences in looking behavior across the different center stimulus types. First, both ASL- and English-learners who processed

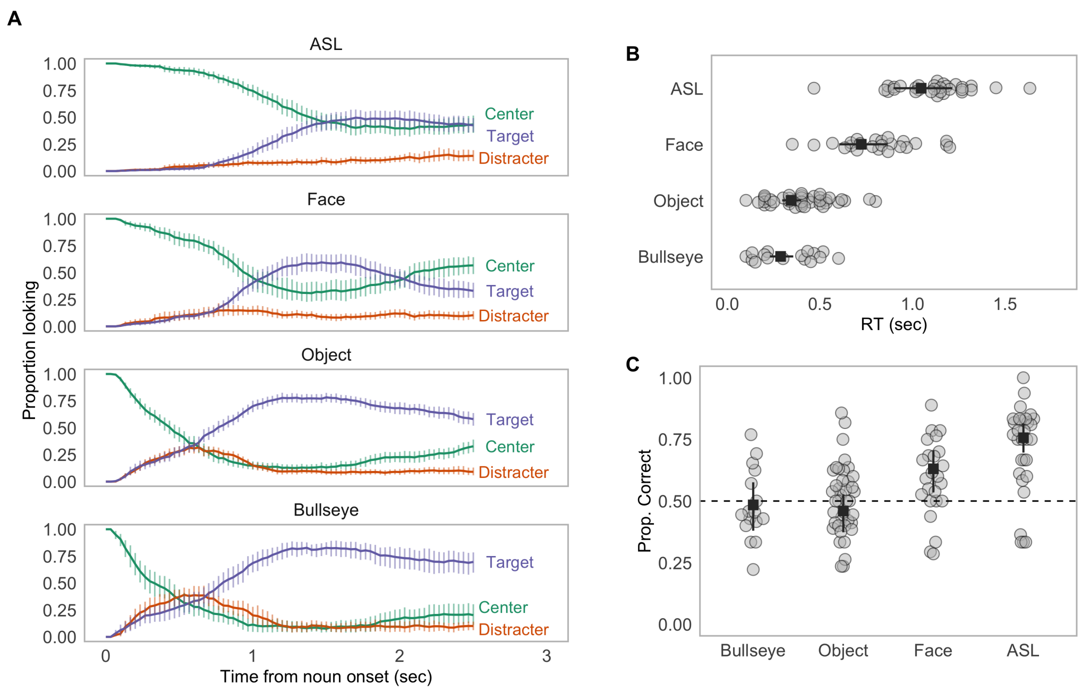

Figure 2. Timecourse looking, first shift Reaction Time (RT), and Accuracy results for children in Experiment 1. Panel A shows the overall looking to the center, target, and distracter stimulus for each context. Panel B shows the distribution of RTs for each participant. Each point represents a participant's average RT. The black squares represent the group means. And the error bars represent 95% Highest Density Intervals around the group means. Panel C shows the same information but for participants' first shift accuracy.

sentences from a video of speaker spent more time looking to the center as indicated by the shallower slope on their center-looking curves. Second, when the center stimulus was a static geometric object (Bullseye) or a static familiar object (Object), spoken language learners were more likely to look at the distracter image, especially early in the timecourse of the target noun as indicated by the parallel increase in target and distracter-looking curves in Figure 2A. In contrast, spoken language learners in the Face context spent less time looking at the disracter, and ASL-learners rarely looked to the distracter image at any point in the trial. This pattern of behavior provides qualitative evidence that children adapated their gaze depending on language-relevant information available in the center stimulus loaction.

Based on a nonparametric cluster-based permutation analysis, the center-looking curve for the ASL learners was significantly different from all other conditions (all $p < .001$). Within the spoken language groups, children's looking to a speaker's face was different from looking to the Bullseye and the Familar object ($p < .001$). Finally, the Object and Bullseye center-looking curves were not different from one another, with no significant differences at any point in the timecourse. Next, we ask how these different processing contexts changed the timing and accuracy of children's initial decisions to shift away from the center stimulus.

RT. Figure 2B shows the full RT data distribution. To quantify differences across the groups, we fit a Bayesian linear mixed-effects regression predicting first shift RT as a function of center stimulus type controlling for age: `Log(RT) ~ center stimulus type + age + (1 | subject) + (1 | item)`. ASL learners generated slower RTs compared to all of the spoken English samples (β = 595.20 ms, 95% HDI [444.60 ms, 760.80 ms]). Moreover, ASL learners' shifts were slower compared directly to children processing spoken language in the Face condition (β = 323.10 ms, 95% HDI [132.30 ms, 522.60 ms]). Finally, children in the Face context shifted gaze slower compared to participants in the Object and Bullseye contexts (β = 408.20 ms, 95% HDI [286.60 ms, 546.20 ms]).

Accuracy. Next, we compared the accuracy of first shifts across the different tasks (Figure 2C) by fitting a mixed-effects logistic regression with the same specifications and contrasts as the RT model. We found that (a) ASL learners were more accurate compared to all of the spoken English samples (β = 0.23, 95% HDI [0.17, 0.29]), (b) ASL learners were more accurate when directly compared to participants in the Face task (β = 0.13, 95% HDI [0.04, 0.23]), (c) children learning spoken language were more accurate when processing language from dynamic video of a person speaking compared to the Object and Bullseye tasks (β = 0.16, 95% HDI [0.07, 0.24]), and (d) English-learners' first shifts were no different from random responding in the Object (β = -0.04, 95% HDI [-0.13, 0.03]) and Bullseye (β = -0.02, 95% HDI [-0.12, 0.08]) contexts.

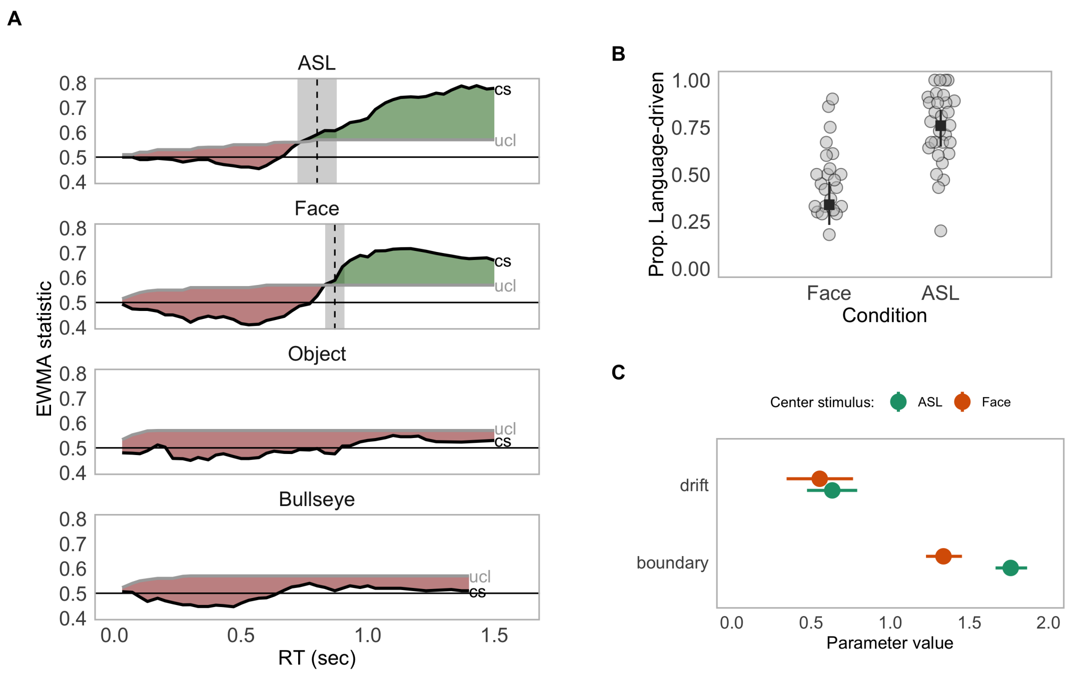

Figure 3. Results for the model-based analyses in Experiment 1. Panel A shows a control chart representing the timecourse of the EWMA statistic. The black curve represents the evolution of the control statistic (CS) as a function of reaction time. The grey curve represents the upper control limit (UCL) or the pre-defined upper limit on random responding. The vertical dashed line is the median cutoff value (point in the RT distribution when children's gaze shifts were no longer random). The grey shaded area represents the 95% Highest Density Interval around the estimate of the median cutoff point, and the shaded ribbons represent the classification of responses as guesses (red, below the UCL) and language-driven (green, above the UCL). Panel B shows a summary of the proportion of shifts that were categorized as language-driven for the Face and ASL processing contexts. Panel C shows the point estimate and 95% Highest Density Intervals for the boundary and drift rate parameters for the Face and ASL contexts.

Model-based analyses. *EWMA.* Our third question of interest was how the tendency to generate random vs. language-driven (i.e, accurate) gaze shifts evolved as a function of reaction time across the different processing contexts. Figure 3A shows changes in the control statistic (CS) and the upper control limit (UCL) as a function of RT. Each CS starts at chance performance and below the UCL. In the ASL and Face tasks, the CS value begins to increase with RTs around 0.7 seconds after noun onset and eventually crosses the UCL, indicating that responses $>$ 0.7 sec were on average above chance levels. In contrast, the CS in the Object and Bullseye tasks never crossed the UCL, indicating that children's shifts were equally likely to land on the target or the distracter, regardless of when they were initiated. This result suggests that first shifts measured in the Bullseye/Object tasks were qualitatively different behaviors than those in the ASL and Face contexts. That is, these shifts are likely the result of a different generative process such as gathering more information about the referents in the visual world.

Next, we compared the EWMA model fits for participants in the ASL and Face processing contexts since these groups showed evidence of language-driven responding. We found that ASL learners generated fewer shifts when the CS was below the UCL compared to children learning spoken language ($\beta = 0.14$, 95% HDI [0.08, 0.23]). This result indicates that ASL-learners were more likely to have gathered sufficient information about the linguistic signal prior to shifting gaze away from the language source. We found some evidence that ASL learners started producing language-driven shifts earlier in the RT distribution as indicated by the point at which the CS crossed the UCL ($\beta = 0.22$ sec, 95% HDI [0.05 sec, 0.39 sec]), indicating that these children were less likely to generate early, random gaze shifts away from the signer.

HDDM. We fit a hierarchical Drift Diffusion Model using only the gaze shifts categorized as language-driven by the EWMA. This allowed us to ask what underlying decision processes are likely to account for the measured differences in First Shift Accuracy and RT.[2] ASL learners had a higher estimate for the boundary separation parameter

[2]We chose not to interpret the HDDM fits for the Bullseye or Face tasks since there was no suggestion of any non-guessing signal from the EWMA analysis.

compared to children processing spoken English from a speaker (ASL boundary = 1.76, 95% HDI [1.65, 1.88]; Face boundary = 1.34, 95% HDI [1.21, 1.47]), with no overlap in the HDIs (see Figure 3C). This suggests that ASL learners' higher accuracy was driven by accumulating more evidence about the linguistic signal before generating an eye movement. We found high overlap for estimates of the drift rate parameter, indicating that both groups processed the linguistic information with similar efficiency (ASL drift = 0.63, 95% HDI [0.44, 0.82]; Face drift = 0.55, 95% HDI [0.30, 0.80]).

Discussion

Taken together, the behavioral and model-based analyses provide converging support that ASL learners were sensitive to the value of delaying eye movements away from the language source. Compared to spoken language learners, ASL learners prioritized accuracy over speed (HDDM), produced fewer nonlanguage-driven shifts away from the center stimulus (EWMA), and were more accurate with these gaze shifts (behavioral). Importantly, we did not see evidence in the HDDM model fits that these accuracy differences could be explained by differential efficiency in processing the linguistic information. Instead, the pattern of results suggests that ASL learners increased their decision threshold to gather more information before shifting gaze away from the signer and to a named object.

We hypothesized that prioritizing accuracy of gaze shifts above speed of responding when processing a visual-manual language is an adaptive response. That is, to map referential language to the visual world in ASL involves competition for visual attention. When ASL learners choose to shift their gaze away from a signer, they are leaving an area that provides a great deal of useful information. Moreover, unlike children learning spoken languages, ASL learners cannot gather more of the linguistic signal if their gaze is directed away from a signer. Thus, it seems reasonable that ASL learners would adapt the timing of their gaze shifts to gather additional information that increases certainty in comprehension before seeking a named object.

It is important to point out that these findings were based on exploratory analyses, and our information seeking explanation was developed to explain this pattern of results. There are, however, several, potentially important differences between the stimuli, apparatus, and populations that limit the strength of our interpretation and the generality of our account. We also cannot make causal conclusions because of the observational nature of research comparing children who are learning different languages. Thus, we designed Experiment 2 to address these concerns and to perform a well-controlled, experimental test of our information seeking explanation of eye movements during grounded language comprehension.

Experiment 2

Experiment 2 provides a well-controlled test of our information seeking explanation of the differences between young ASL- and English-learners that we found in Experiment 1.[3] We measured adults and children's eye movements during a real-time language comprehension task where participants processed familiar sentences (e.g., "Where's the ball?") while looking at a simplified visual world with three fixation targets. Using a within-participants design, we manipulated the signal-to-noise ratio of the auditory signal by convolving the acoustic input with brown noise (random noise with greater energy at lower frequencies). We chose a noise manipulation because it allowed us to increase the value of looking to a speaker for language comprehension, and it could be used with both adults and children.

We predicted that processing speech in a noisy context would make participants less likely to shift before collecting sufficient information. This delay, in turn, would lead to a lower proportion of shifts flagged as random/exploratory in the EWMA analysis. We also

[3] See https://osf.io/g8h9r/ for a pre-registration of the analysis plan.

predicted a developmental difference: that children would produce a higher proportion of random shifts and accumulate information less efficiently compared to adults, and a developmental parallel: that children would show similar patterns of adapting gaze to gather additional visual information in the noisier auditory environment.

Methods

Participants. Participants were native, monolingual English-learning children (n = 39; 22 F) and adults (n = 31; 22 F). All participants had no reported history of developmental or language delay and normal vision. 14 participants (11 children, 3 adults) were run but not included in the analysis because either the eye tracker falied to calibrate (2 children, 3 adults) or the participant did not complete the task (9 children).

Stimuli. *Linguistic stimuli.* The video/audio stimuli were recorded in a sound-proof room and featured two female speakers who used natural child-directed speech and said one of two phrases: “Hey! Can you find the (target word)” or "Look! Where’s the (target word). The target words were: ball, bunny, boat, bottle, cookie, juice, chicken, and shoe. The target words varied in length (shortest = 411.68 ms, longest = 779.62 ms) with an average length of 586.71 ms.

Noise manipulation. To create the stimuli in the noise condition, we convolved each recording with Brown noise using the Audacity audio editor. The average signal-to-noise ratio (values greater than 0 dB indicate more signal than noise) in the noise condition was 2.87 dB compared to the clear condition, which was 35.05 dB.

Visual stimuli. The image set consisted of colorful digitized pictures of objects presented in fixed pairs with no phonological overlap between the target and the distracter image (cookie-bottle, boat-juice, bunny-chicken, shoe-ball). The side of the target picture was counterbalanced across trials.

Design and procedure. Participants viewed the task on a screen while their gaze was tracked using an SMI RED corneal-reflection eye-tracker mounted on an LCD monitor, sampling at 30 Hz. The eye-tracker was first calibrated for each participant using a 6-point calibration. On each trial, participants saw two images of familiar objects on the screen for two seconds before the center stimulus appeared. Next, they processed the target sentence – which consisted of a carrier phrase, a target noun, and a question – followed by two seconds without language to allow for a response. Child participants saw 32 trials (16 noise trials; 16 clear trials) with several filler trials interspersed to maintain interest. Adult participants saw 64 trials (32 noise; 32 clear). The noise manipulation was presented in a blocked design with the order of block counterbalanced across participants.

Results and Discussion

Behavioral analyses. *Timecourse looking.* Figure 4A presents an overview of looking to the speaker, target, and distracter images for the noisy and clear processing contexts from the start of the target noun. Similar to the results in Experiment 1, participants tended to fixate on the speaker at target-noun onset. As the target noun unfolded, the mean proportion looking to the center decreased rapidly as participants shifted their gaze to the objects. Proportion looking to the target increased sooner and reached a higher asymptote compared to proportion looking to the distracter for both processing contexts and age groups. After looking to the target image, participants tended to shift their gaze back to the speaker as shown by the increase in center looking curve around 1 second.

There are several developmental differences to highlight. First, children tended to look more to the objects at noun onset, as indicated by the lower intercept of children's center-looking curves. Second, children's target looking curves reached a lower asymptote as compared to adults and they spent relatively more time fixating on the distracter image, whereas adults rarely looked at the unnamed object after 0.5 seconds in the timecourse of

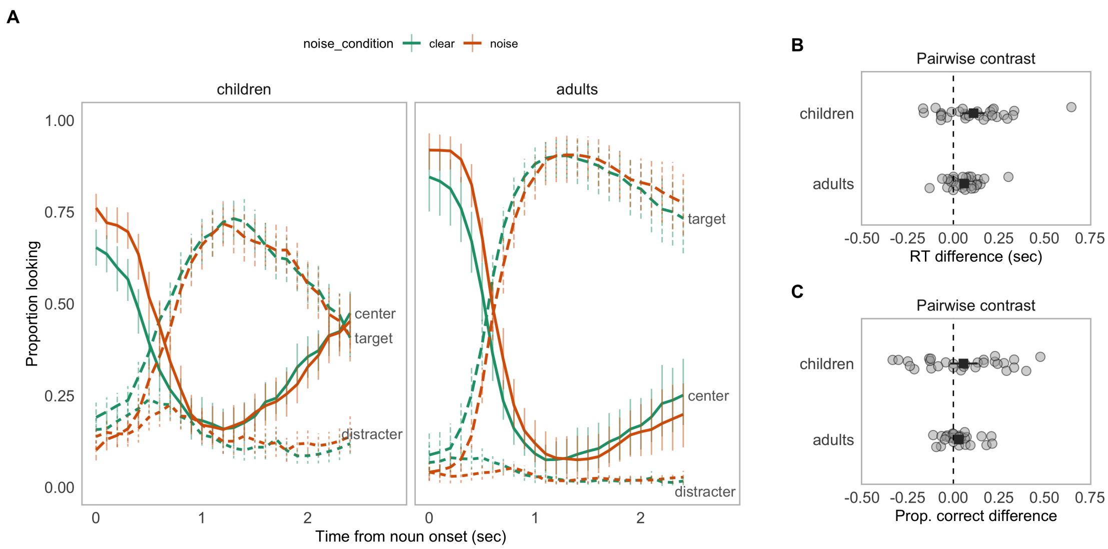

Figure 4. Behavioral results for children and adults in Experiment 2. Panel A shows the overall looking to the center, target, and distracter stimulus for each processing condition and age group. Panel B shows the distribution of RTs for each participant and the pairwise contrast between the noise and clear conditions. The square point represents the mean value for each mesure. The vertical dashed line represents the null model of zero condition difference. The width each point represents the 95% HDI. Panel C shows the same information but for first shift accuracy.

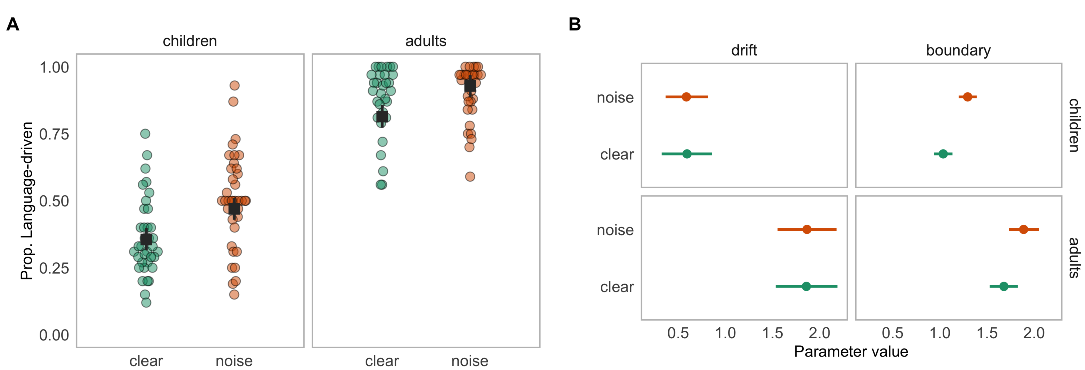

Figure 5. Results for the model-based analyses for Experiment 2. The plotting conventions are the same as Figure 3.

the trial. And third, children showed a stronger tendency to shift back to the speaker after looking to the named object.

Visual inspection of the center looking curves suggests a difference in looking behavior in the noisy processing context. Both children and adult's spent more time fixating on the speaker when the auditory signal was less reliable as indicated by the rightward shift of the center-looking curves in the noisy condition. A cluster-based permutation test confirmed that there was evidence of a significant difference in looking to the speaker between the Noisy and Clear conditions ($p < .05$). This pattern of behavior provides evidence that reducing the quality of the auditory signal increased looking to the speaker early in the timecourse of the target noun.

RT. Figure 4B shows the full distribution of the estimated RT differences between each participants' performance in the noisy and clear contexts. Both children and adults were slower to identify the target in the noise condition (Children M_{noise} = 500.20 ms; Adult M_{noise} = 595.20 ms), as compared to the clear condition (Children M_{clear} = 455.70 ms Adult M_{clear} = 542.40 ms). RTs in the noise condition were 48.80 ms slower on average, with a 95% HDI ranging from 3.70 ms to 96.30 ms, and not including the null value of zero condition difference. Older children responded faster than younger children (β_{age} = -0.44, [-0.74, -0.16]), with little evidence for an interaction between age and condition within the child sample.

Accuracy. Next, we modeled adults and children's first shift accuracy using a mixed-effects logistic regression with the same specifications (Figure 4C). Both groups were more accurate than a model of random responding with the null value of 0.5 falling well outside the lower bound of the 95% HDI for each group mean. Adults were more accurate (M_{adults} = 90%) than children ($M_{children}$ = 61%). Interestingly, both groups showed evidence of higher accuracy in the noise condition: children (M_{noise} = 67%; M_{clear} = 61%) and adults (M_{noise} = 92%; M_{clear} = 90%). Accuracy in the noise condition was on average 4% higher, with a 95% HDI from -1% to 12%. Note that the null value of zero difference falls at the very edge of the HDI. But 95% of the credible values are greater than zero, providing evidence for comparable, if not higher, accuracy in the noise condition. Within the child sample, there was no evidence of a main effect of age or an interaction between age and noise condition.

Model-based analyses. **EWMA.** Figure 5A shows the proportion of shifts that the model classified as random vs. language-driven for each age group and processing context. On average, 41% (95% HDI: 32%, 50%) of children's shifts were categorized as language-driven, which was significantly fewer than adults, 87% (95% HDI: 78%, 96%). Critically, processing speech in a noisy context caused both adults and children to generate a higher proportion of language-driven shifts (i.e., fewer random, exploratory shifts away from the speaker), with the 95% HDI excluding the null value of zero condition difference ($\beta_{noise} =$ 11%, [7.00%, 16%]). Within the child sample, older children generated fewer random, early shifts ($M_{age} =$ -0.21, [-0.35, -0.08]). There was no eivdence of an interaction between age and condition. This pattern of results suggests that the noise condition caused participants to increase visual fixations to the language source, leading them to generate fewer exploratory, random shifts before they had accumulated sufficient information to respond accurately.

HDDM. Figure 5B shows the full posterior distributions for the HDDM output. Children had lower estimats of drift rate (children $M_{drift} = 0.59$; adults $M_{drift} = 1.90$) and boundary separation (children $M_{boundary} = 1.16$; adults $M_{boundary} = 1.67$) as compared to adults, suggesting that children were both less efficient and less cautious in their shifts to the named object. The noise manipulation selectively affected the boundary separation parameter, with higher estimates in the noise condition for both age groups ($\beta_{noise} = 0.26$, [0.10, 0.42]). This result suggests that participants in the noise condition prioritized information accumulation over speed when generating an eye movement in response to the incoming language, and this increased decision threshold led to higher accuracy. Moreover, the high overlap in estimates of drift rate suggests that participants were able to integrate the visual and auditory signals such that they could achieve a level of processing efficiency comparable to the clear processing context.

Taken together, the behavioral and EWMA/HDDM results provide evidence for our information-seeking explanation of eye movements during grounded language comprehension. Processing speech in noise caused both children and adults to look longer at their social partner, which in turn, resulted in a higher proportion of accurate gaze shifts to a named object. Moreover, we observed a similar pattern of behavior in children and adults, with both groups producing more language-driven shifts (EWMA) and prioritizing accuracy over speed (HDDM) in the more challenging noisy environment.

It is important to point out the correspondence between the reported results and our pre-registered analysis plan and predictions. First, we predicted that the noise manipulation would cause listeners to gather more information by looking longer at the speaker's face (slower RTs), and that this behavior would lead listeners to produce a higher proportion of language-driven shifts as indexed by the EWMA analysis. We did not, however, predict that first shifts would be *more* accurate in the noisier context and that the noise manipulation would selectively affect the boundary separation parameter in the HDDM. Finally, we chose not to collect our planned sample size (42) for each age group (3-, 4-, and 5-year-olds) in the target age range; instead we collected a single sample that included children across the entire age range.

General Discussion

Language comprehension in grounded, social contexts provides children access to a rich set of multimodal cues that could support the linking of linguistic information to the world. But how do children select what information to gather? In this work, we proposed that listeners flexibly adapt their gaze to seek visual information from their social partners when it was especially useful for language comprehension. We presented evidence for our account by measuring changes in how children chose to allocate visual attention across two diverse language processing contexts. In Experiment 1, we found that, compared to children learning spoken English, young ASL-learners delayed their gaze shifts away from a language source, were more accurate, and produced a higher proportion of language-driven eye movements. In Experiment 2, we showed that 3-5 year-olds and adults delayed the timing of gaze shifts away from a speaker's face while processing speech in a noisy auditory environment. This slower response resulted in fewer nonlanguage-driven eye movements and more accurate gaze shifts.

These results synthesize ideas from several research programs, including work on language-driven visual attention (Tanenhaus et al., 1995), goal-based accounts of vision during everyday tasks (Hayhoe & Ballard, 2005), and work on language perception as multisensory integration (Vigliocco et al., 2014). Moreover, our findings parallel the results of several recent studies that measure the adaption of visual processes in response to different auditory experiences. First, Heimler et al. (2015) compared Deaf and hearing adults' performance on an oculomotor singleton detection paradigm where participants made speeded eye-movements to a unique target embedded among distracters that varied in saliency. Deaf adults were slower to generate a gaze shift away from the center fixation and, as a result, they were less affected by high saliency distracters. Second, McMurray, Farris-Trimble, and Rigler (2017) found that individuals with cochlear implants, who are consistently processing degraded auditory input, are more likely to delay the process of lexical access as measured by slower gaze shifts to named referents and fewer incorrect gaze shifts to phonological onset competitors. McMurray et al. (2017) also found that they could replicate these changes in adults with typical hearing by noise-vocoded speech stimuli that shared features with the output of a cochlear implant.

Our findings also connect to the literature investigating how experience with a visual-manual language may change basic cognitive processes (see Bavelier, Dye, and Hauser (2006) for a review). The upshot of this work is that the effects of Deafness can be dissociated from the effects of learning a signed language. Specifically, Deaf individuals show selective enhancement in peripheral visual attention as evidenced by higher sensitivity to peripheral distracters on spatial orienting tasks. In contrast, learning to sign results in several specific changes such as enhanced mental imagery (Emmorey, Kosslyn, & Bellugi, 1993), mental rotation (Emmorey, Klima, & Hickok, 1998), and face processing (Bettger, Emmorey, McCullough, & Bellugi, 1997). The results of Experiment 1 suggest that ASL learners adapt the timing of when they disengage from a language source to increase their certainty before seeking named object. It is an open question as to whether ASL-learners' differential responding is best explained by lack of access to auditory information or learning a visual-manual language.

Finally, our results dovetail with recent developmental work by Yurovsky et al. (2017). In their study, preschoolers, like adults, were able to integrate top-down expectations about the kinds of things speakers are likely to talk about with bottom-up cues from auditory perception. Yurovsky et al. (2017) situated this finding within the framework of modeling language as a *noisy channel* where listeners combine expectations with perceptual data and weight each based on its reliability. In Experiment 2, we found a similar developmental parallel in language processing: that 3-5 year-olds, like adults, adapted their gaze patterns to seek additional visual information when the auditory signal became less reliable. This adaptation allowed listeners to generate comparable, if not more, accurate responses in the noisy context.

In sum, the work reported here shows that young listeners can seek visual information to support language comprehension. These results fit well with the interactive models of language perception reviewed in the Introduction (MacDonald & Seidenberg, 2006; McClelland et al., 2006). These studies also also highlight the value of using both observational and experimental approaches. In Experiment 1, we compared language comprehension across populations of children who had very different language experiences (signed vs. spoken) to generate a novel explanation of observed differences in children's gaze dynamics. We were able to better understand this observational result by designing a well-controlled, follow-up experiment that tested predictions of our explanation and allowed us to make stronger claims about the generality of our hypothesis.

Limitations and future work

Our results provide evidence that young listeners can adapt their gaze patterns to the demands of different processing environments to seek visual information from social partners that supports language comprehension. We cannot, however, make claims about how children's behavior in our task (the Visual World Paradigm: VWP) would generalize to their decisions about how to distribute attention in real-world learning environments. There is a growing body of research showing meaningful links between children's gaze behavior in the VWP and relevant outcome measures such as vocabulary development (Fernald et al., 2006; Marchman & Fernald, 2008; Rigler et al., 2015). Nonetheless, a valuable next step for our work would be to leverage tasks and measures that are come closer to the ecological context in which children actually processs and learn language such as using head-mounted cameras and eye trackers that would allow measurement of where children choose to look during everyday interactions (Fausey, Jayaraman, & Smith, 2016; Franchak, Kretch, Soska, & Adolph, 2011).

This work has several other important limitations. First, we chose to focus on a single decision about visual fixation to provide a window onto the dynamics of decision-making across different language processing contexts. But our analysis does not consider the rich information present in the gaze patterns that occur leading up to this decision. In our future work, we aim to measure how changes in the language environment might lead to shifts in the dynamics of gaze across a longer timescale. For example, perhaps listeners gather more information about the objects in the scene before the sentence in anticipation of allocating more attention to the speaker once they start to speak.

Second, we chose one instantiation of a noisy processing context – random background noise. But we think our findings should generalize to contexts where other kinds of noise – e.g., uncertainty over a speaker's reliability or when processing accented speech – make gathering visual information from the speaker more useful for language understanding. Moreover, we used a simple visual world, with only three places to look, and simple linguistic stimuli. Thus it remains an open question how these results might scale up to more complex language interactions and visual environments. It could be that looks to a speaker become even more useful for disambiguating reference in complex visual environments.

Third, we do not yet know what might be driving the population differences between children learning ASL and children learning spoken English found in Experiment 1. It could be that ASL-learners' massive experience dealing with competition for visual attention leads to changes in the deployment of eye movements during language comprehension. Or, it could be that the in-the-moment constraints of processing a visual language cause different fixation behaviors. This question could be addressed by studies that measure how quickly listeners adapt the dynamics of gaze when visual information becomes more useful. Another interesting approach would be to measure eye movements in hearing children learning both a signed and a spoken language (bimodal bilinguals). This comparison between native hearing and deaf signers would allow for a dissociation of the effects of learning a visual-manual language from the effects of lacking access to auditory information (e.g., Bavelier et al., 2006). If hearing signers also prioritize accuracy over speed when processing their spoken language, this would suggest that experience with a visual-manual language is changing a general response strategy.

Finally, our eye tracking paradigm removes an important component of successful communication: dynamic interaction between the speaker and listener. It is interesting to consider how speakers might adapt their behavior present the listener with useful visual information in challenging comprehension contexts. For example, in noisy environments, speakers will exaggerate mouth movements (Fitzpatrick, Kim, & Davis, 2011) and increase the frequency of gestural cues such as head nodding (Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004), and parents exaggerate mouth movements during infant-directed speech (Green, Nip, Wilson, Mefferd, & Yunusova, 2010). Moreover, observational studies of parent-child interactions in signed languages show variability in how sensitive adult signers are to the competing demands on children's visual attention (Harris & Mohay, 1997). That is, some interactions contain many utterances that young signers miss because they are fixating on objects; whereas other interactions are marked by adaptations that accommodate the demands on visual attention by parents displacing signs onto the objects that are currently the focus of children's attention (similar to follow-in labeling effects Tomasello and Farrar, 1986). Thus it is an interesting, open question how interacting with a speaker that adapts to increase the availability and utility of visual information might change children's decisions about visual fixation.

Conclusion

In this paper, we presented an information-seeking explanation for the differences in the dynamics of eye movements during grounded signed vs. spoken language comprehension. We started from an interesting, observational result: that ASL learners, compared to English-learning children, generate slower but more accurate gaze shifts away from a language source and to a named referent. We then tested the generality and causal claims of our account by experimentally manipulating the value of visual information for language comprehension. We found that young listeners can adapt the dynamics of their gaze to gather visual information when it is useful for language understanding.

While we chose to start with the domain of familiar language processing, we think the account is more general and could could be applied to the language acquisition context.

Consider that early in language learning, children are acquiring novel word-object links while also learning about visual object categories. Both of these tasks produce different goals that should, in turn, modulate children's decisions about where to allocate visual attention – e.g., seeking nonlinguistic cues to reference such as eye gaze and pointing become critical when you are unfamiliar with the information in the linguistic signal. More generally, we think that this approach presents a way forward for explaining decisions about visual fixation during language comprehension and acquisition across a larger set of processing contexts and at different stages of development.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419–439.

Altmann, G., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, *57*(4), 502–518.

Bavelier, D., Dye, M. W., & Hauser, P. C. (2006). Do deaf individuals see better? *Trends in Cognitive Sciences*, *10*(11), 512–518.

Bettger, J. G., Emmorey, K., McCullough, S. H., & Bellugi, U. (1997). Enhanced facial discrimination: Effects of experience with american sign language. *Journal of Deaf Studies and Deaf Education*, 223–233.

Clark, E. V. (2009). *First language acquisition*. Cambridge University Press.

Dahan, D., & Tanenhaus, M. K. (2005). Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word recognition. *Psychonomic*

Bulletin & Review, *12*(3), 453–459.

Emmorey, K., Klima, E., & Hickok, G. (1998). Mental rotation within linguistic and non-linguistic domains in users of american sign language. *Cognition*, *68*(3), 221–246.

Emmorey, K., Kosslyn, S. M., & Bellugi, U. (1993). Visual imagery and visual-spatial language: Enhanced imagery abilities in deaf and hearing asl signers. *Cognition*, *46*(2), 139–181.

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research*, *12*(2), 423–425.

Estigarribia, B., & Clark, E. V. (2007). Getting and maintaining attention in talk to young children. *Journal of Child Language*, *34*(4), 799–814.

Fausey, C. M., Jayaraman, S., & Smith, L. B. (2016). From faces to hands: Changing visual input in the first two years. *Cognition*, *152*, 101–107.

Fernald, A., Perfors, A., & Marchman, V. A. (2006). Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. *Developmental Psychology*, *42*(1), 98.

Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language. *Developmental Psycholinguistics: On-Line Methods in Children's Language Processing*, *44*, 97.

Fitzpatrick, M., Kim, J., & Davis, C. (2011). The effect of seeing the interlocutor on auditory and visual speech production in noise. In *Auditory-visual speech processing 2011.*

Fourtassi, A., & Frank, M. C. (2017). Word identification under multidomodal uncertainty. In *Proceedings of the 39th annual conference of the cognitive science society.*

Franchak, J. M., Kretch, K. S., Soska, K. C., & Adolph, K. E. (2011). Head-mounted eye tracking: A new method to describe infant looking. *Child Development*, *82*(6), 1738–1750.

Gabry, J., & Goodrich, B. (2016). Rstanarm: Bayesian applied regression modeling via stan. R package version 2.10. 0.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 201216438.

Gold, J. I., & Shadlen, M. N. (2000). Representation of a perceptual decision in developing oculomotor commands. *Nature*, *404*(6776), 390.

Green, J. R., Nip, I. S., Wilson, E. M., Mefferd, A. S., & Yunusova, Y. (2010). Lip movement exaggerations during infant-directed speech. *Journal of Speech, Language, and Hearing Research*, *53*(6), 1529–1542.

Harris, M., & Mohay, H. (1997). Learning to look in the right place: A comparison of attentional behavior in deaf children with deaf and hearing mothers. *The Journal of Deaf Studies and Deaf Education*, *2*(2), 95–103.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, *9*(4), 188–194.

Heimler, B., Zoest, W. van, Baruffaldi, F., Donk, M., Rinaldi, P., Caselli, M. C., & Pavani, F. (2015). Finding the balance between capture and control: Oculomotor selection in early deaf adults. *Brain and Cognition*, *96*, 12–27.

Hoppe, D., & Rothkopf, C. A. (2016). Learning rational temporal eye movement strategies. *Proceedings of the National Academy of Sciences*, *113*(29), 8332–8337.

Huettig, F., & Altmann, G. T. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. *Cognition*, *96*(1), B23–B32.

Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, *21*(2), 260–267.

Liszkowski, U., Brown, P., Callaghan, T., Takada, A., & De Vos, C. (2012). A prelinguistic gestural universal of human communication. *Cognitive Science*, *36*(4), 698–713.

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Attention, Perception, & Psychophysics*, *24*(3), 253–257.

MacDonald, K., LaMarr, T., Corina, D., Marchman, V. A., & Fernald, A. (2018). Real-time lexical comprehension in young children learning american sign language. *Developmental Science*, e12672.

MacDonald, M. C., & Seidenberg, M. S. (2006). Constraint satisfaction accounts of lexical and sentence comprehension. *Handbook of Psycholinguistics*, *2*, 581–611.

Marchman, V. A., & Fernald, A. (2008). Speed of word recognition and vocabulary knowledge in infancy predict cognitive and language outcomes in later childhood. *Developmental Science*, *11*(3).

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of eeg-and meg-data. *Journal of Neuroscience Methods*, *164*(1), 177–190.

McClelland, J. L., & Elman, J. L. (1986). The trace model of speech perception. *Cognitive Psychology*, *18*(1), 1–86.

McClelland, J. L., Mirman, D., & Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Sciences*, *10*(8), 363–369.

McMurray, B., Farris-Trimble, A., & Rigler, H. (2017). Waiting for lexical access: Cochlear implants or severely degraded input lead listeners to process speech less incrementally. *Cognition*, *169*, 147–164.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science*, *15*(2), 133–137.

Nelson, J. D., & Cottrell, G. W. (2007). A probabilistic model of eye movements in concept formation. *Neurocomputing*, *70*(13-15), 2256–2272.

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, *68*, 169–181.

Ratcliff, R., & Childers, R. (2015). Individual differences and fitting methods for the two-choice diffusion model of decision making. *Decision*, *2*(4), 237–279.

Rigler, H., Farris-Trimble, A., Greiner, L., Walker, J., Tomblin, J. B., & McMurray, B. (2015). The slow developmental time course of real-time spoken word recognition. *Developmental Psychology*, *51*(12), 1690.

Salverda, A. P., Brown, M., & Tanenhaus, M. K. (2011). A goal-based perspective on eye movements in visual world studies. *Acta Psychologica*, *137*(2), 172–180.

Shinoda, H., Hayhoe, M. M., & Shrivastava, A. (2001). What controls attention in natural environments? *Vision Research*, *41*(25-26), 3535–3545.

Spivey, M. J., Tanenhaus, M. K., Eberhard, K. M., & Sedivy, J. C. (2002). Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*, *45*(4), 447–481.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632.

Tomasello, M., & Farrar, M. J. (1986). Joint attention and early language. *Child Development*, 1454–1463.

Triesch, J., Ballard, D. H., Hayhoe, M. M., & Sullivan, B. T. (2003). What you see is what you need. *Journal of Vision*, *3*(1), 9–9.

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review*, *14*(6), 1011–1026.

Venker, C. E., Eernisse, E. R., Saffran, J. R., & Weismer, S. E. (2013). Individual differences in the real-time comprehension of children with asd. *Autism Research*, *6*(5), 417–432.

Vigliocco, G., Perniss, P., & Vinson, D. (2014). Language as a multimodal phenomenon: Implications for language learning, processing and evolution. The Royal Society.

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical bayesian estimation of the drift-diffusion model in python. *Frontiers in Neuroinformatics*, *7*, 14.

Yee, E., & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(1), 1.

Yurovsky, D., Case, S., & Frank, M. C. (2017). Preschoolers flexibly adapt to linguistic input in a noisy channel. *Psychological Science*, *28*(1), 132–140.

Appendix A

Experiment 1

Table A1

Output of the regression predicting reaction time (milliseconds) as a function of center stimulus type in Experiment 1.

Center Stimulus Type	Mean RT	95% HDI
Bullseye	288.71	[229.8, 355.17]
Object	344.51	[295.71, 397.31]
Face	724.78	[603.2, 862.81]
ASL	1,047.89	[897.29, 1213.52]

Table A2

Output of the logistic regression predicting accuracy as a function of center stimulus type in Experiment 1.

Center Stimulus Type	Mean Accuracy	95% HDI
Object	0.46	[0.37, 0.53]
Bullseye	0.48	[0.38, 0.58]
Face	0.63	[0.54, 0.7]
ASL	0.76	[0.7, 0.81]

Table A3

Output of the model estimating differences in Accuracy for specific contrasts of interest in Experiment 1.

Contrast	Mean Difference Accuracy	95% HDI
Object vs. Chance	-0.04	[-0.13, 0.03]
Bullseye vs. Chance	-0.02	[-0.12, 0.08]
ASL vs. Face	0.13	[0.04, 0.23]
Face vs. Object/Bullseye	0.16	[0.07, 0.24]
ASL vs. English	0.23	[0.17, 0.29]

Table A4

Output of the model estimating differences in RT for specific contrasts of interest in Experiment 1.

Contrast	Mean Difference RT	95% HDI
ASL vs. Face	323.10	[132.3, 522.6]
Face vs. Object/Bullseye	408.20	[286.6, 546.2]
ASL vs. English	595.20	[444.6, 760.8]

Table A5

Output of the model estimating the point in the Reaction Time distribution when children's Exponentially Weighted Moving Average statistic crossed the pre-defined guessing threshold for the ASL and Face center stimulus types in Experiment 1.

Center Stimulus Type	Mean EWMA Cut Point	95% HDI
ASL	0.68	[0.59, 0.78]
Face	0.90	[0.77, 1.03]

Table A6

Output of the model estimating the mean proportion of shifts categorized as language-driven by the Exponentially Weighted Moving Average model for the ASL and Face center stimulus types in Experiment 1.

Center Stimulus Type	Mean Language-driven	95% HDI
Face	0.34	[0.23, 0.46]
Asl	0.75	[0.65, 0.84]

Table A7

Summary of the Drift Diffusion Model output for the drift rate and boundary separation parameters for both all four center stimulus types in Experiment 1.

Parameter	Center Stim Type	Mean Param Estimate	95% HDI
Boundary	Face	1.34	[1.21, 1.47]
Boundary	ASL	1.76	[1.65, 1.88]
Drift	Face	0.55	[0.3, 0.8]
Drift	ASL	0.63	[0.44, 0.82]

Appendix B

Experiment 2

Table B1

Output of the logistic regression predicting accuracy as a function of noise condition and age group in Experiment 2.

Noise Condition	Age Group	Mean Accuracy	95% HDI
Clear	children	0.61	[0.54, 0.68]
Noise	children	0.67	[0.6, 0.74]
Clear	adults	0.90	[0.87, 0.93]
Noise	adults	0.92	[0.89, 0.95]

Table B2

Output of the regression estimating reaction times (milliseconds) as a function of noise condition and age group in Experiment 2.

Noise Condition	Age Group	Mean RT	95% HDI
Clear	children	455.70	[407, 503.6]
Noise	children	500.20	[446.6, 555.6]
Clear	adults	542.40	[486.2, 602.4]
Noise	adults	595.20	[532.4, 665.3]

Table B3

Output of the model estimating the mean proportion of shifts categorized as language-driven by the Exponentially Weighted Moving Average model for the each noise condition and age group in Experiment 2.

Noise Condition	Age Group	Mean Language-driven	95% HDI
Clear	children	0.36	[0.32, 0.4]
Noise	children	0.47	[0.43, 0.51]
Clear	adults	0.81	[0.77, 0.86]
Noise	adults	0.93	[0.89, 0.97]

Table B4

Summary of the Drift Diffusion Model output for the drift rate and boundary separation parameters for both processing contexts and age groups in Experiment 2.

Parameter	Age Group	Mean Parameter Estimate	95% HDI
Boundary	Children	1.16	[0.94, 1.39]
Boundary	Adults	1.67	[1.49, 1.84]
Drift	Children	0.59	[0.3, 0.89]
Drift	Adults	1.90	[1.51, 2.3]