1   An information-seeking account of children's eye movements during grounded signed and

2   spoken language comprehension

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

4   [1] Stanford University

5   Author Note

11   Abstract

12   Language comprehension in grounded, social interactions involves extracting meaning from

13   the linguistic signal and mapping it to the visual world. Information that is gathered

14   through visual fixations can facilitate this comprehension process. But how do listeners

15   decide what visual information to gather and at what time? Here, we propose that listeners

16   flexibly adapt their gaze behaviors to seek visual information from their social partners to

17   support robust language understanding. We present evidence for our account using three

18   case studies of eye movements during real-time language processing. First, compared to

19   children (n=80) and adults (n=25) learning spoken English, young ASL-learners (n=30) and

20   adults (n= 16) delayed gaze shifts away from a language source, were more accurate with

21   these shifts, and produced a smaller proportion of random shifts. Next, English-speaking

22   adults (n=30) produced fewer random gaze shifts when processing serially printed text

23   compared to processing spoken language. Finally, English-speaking preschoolers (n=39) and

24   adults (n=31) delayed the timing of shifts away from a speaker while processing speech in

25   noisy environments, gathering more visual information while generating more accurate

26   responses. Together, these results provide evidence that listeners adapt to the demands of

27   different processing contexts by seeking out visual information from social partners.

28   *Keywords:* eye movements; language comprehension; information-seeking; speech in

29   background noise; American Sign Language

30   Word count: X

31   An information-seeking account of children's eye movements during grounded signed and

32   spoken language comprehension

Introduction

34   Extracting meaning from language represents a formidable challenge for young

35   language learners. Consider that even in the simple case of understanding grounded, familiar

36   language (e.g., "look at the ball"), the listener must continuously integrate linguistic and

37   non-linguistic information from continuous streams of input. Moreover, language unfolds

38   within dynamic interactions where there is often insufficient information to figure out what is

39   being said, and yet the listener must decide how best to respond. Even young children,

40   however, can map language to the world quite efficiently, shifting visual attention to a named

41   object in a scene within hundreds of milliseconds upon hearing the name of an object

42   (Allopenna, Magnuson, & Tanenhaus, 1998; Spivey, Tanenhaus, Eberhard, & Sedivy, 2002;

43   Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). How do young listeners

44   successfully interpret linguistic input despite these challenges?

45   One solution is for the language comprehension system to integrate multiple sources of

46   information to constrain the set of possible interpretations (MacDonald & Seidenberg, 2006;

47   McClelland & Elman, 1986). Under this interactive account, listeners comprehend words by

48   partially activating several candidates that are consistent with the incoming perceptual

49   information. Then, as more information arrives, words that do not match the perceptual

50   signal are no longer considered, and words that are more consistent become strongly

51   activated until a single interpretation is reached (see McClelland, Mirman, and Holt (2006)

52   for a review). Critically, information from multiple sources – e.g., the linguistic signal and

53   visual world – mutually influence one another to shape interpretation. For example, if a

54   speaker's mouth movements suggest one sound while their acoustic output indicates another,

55   the conflict results in the listener perceiving a third, intermediate sound ("McGurk effect")

56   (MacDonald & McGurk, 1978). Other research shows that listeners will use information in

57   the visual scene to help parse syntactically ambiguous utterances (Tanenhaus et al., 1995).

58   Thus, information gathered from the visual world can serve as a useful constraint on

59 language comprehension. But the incoming linguistic information is ephemeral, meaning

60 listeners must quickly decide how to direct their gaze to informative locations. Consider a

61 speaker who asks you to "Pass the salt" in a noisy restaurant. Here, comprehension could be

62 supported by looks that better encode the objects in the scene (e.g., the type of food she is

63 eating), or by looks to the speaker (e.g., reading her lips or the direction of her gaze). A

64 second interesting case is the processing a visual-manual language such as American Sign

65 Language (ASL). Here, deciding to look away from another signer is inherently risky because

66 the listener stops the flow of information from the linguistic signal to gather information

67 about the visual world.

68   Eye movements during language comprehension can be characterized an active

69 decision-making process taking into account limits on visual attention. We propose that

70 listeners are sensitive to this tradeoff, flexibly adapting the dynamics of their gaze in

71 contexts that place a higher value on gathering visual information. That is, we suggest that

72 listeners' eye movements are shaped by an interaction between their sensorimotor constraints

73 and information features of the environment.

74   Our account is inspired by ideas from several research traditions. First, work on

75 language-mediated visual attention showing rapid interactions between visual attention and

76 language (Allopenna et al., 1998; Tanenhaus et al., 1995). Second, research on vision in

77 everyday tasks shows that people allocate fixations to *goal-relevant* locations – e.g., an

78 upcoming obstacle while walking (Hayhoe & Ballard, 2005). Finally, work on multisensory

79 integration showing that listeners leverage multimodal cues (e.g., gestures, facial expressions,

80 mouth movements) to support communication. In the following sections, we review each of

81 these literatures to motivate our information-seeking account of eye movements in social,

82 grounded language comprehension.

Vision-Language interactions during language comprehension

Eye movements during language comprehension have 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 people will initiate gaze shifts to naemd 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 provides 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 tended to look at phonological onset-competitor (a beetle) early in the target noun, suggesting that they had activated multiple interpretations and resolved ambiguity as the stimulus unfolded. These behavioral results fall out of predictions made by interactive models of speech perception where information from multiple sources is integrated to constrain language understanding (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 propose that anticipatory eye movements reflect the influence of the visual information in a scene activating a multi-featured, conceptual representation prior to the arrival of the linguistic signal (see also Huettig and Altmann (2005)).

In addition to work on adult psycholinguistics, the VWP has been useful for studying developmental change in language comprehension skill in children. Researchers have adapted

110   the task to measure the timing and accuracy of children's gaze shifts as they look at two

111   familiar objects and listen to simple sentences naming one of the objects (Fernald, Zangl,

112   Portillo, & Marchman, 2008; Venker, Eernisse, Saffran, & Weismer, 2013). Such research

113   finds that children, like adults, shift gaze to named objects occur soon after the auditory

114   information is sufficient to enable referent identification. Moreover, individual differences in

115   the speed and accuracy of eye movements predict vocabulary growth and later language and

116   cognitive outcomes (Fernald, Perfors, & Marchman, 2006; Marchman & Fernald, 2008; Rigler

117   et al., 2015). Finally, the VWP illustrated interesting developmental parallels and differences

118   between children's language processing in different populations, including sign language

119   (MacDonald, LaMarr, Corina, Marchman, & Fernald, 2018), bilingualism (Byers-Heinlein,

120   Morin-Lessard, & Lew-Williams, 2017), and children with cochlear implants (Schwartz,

121   Steinman, Ying, Mystal, & Houston, 2013).

Goal-based accounts of eye movements in everyday tasks

123   The majority of the work on language-mediated visual attention has used eye

124   movements as an index of the underlying interaction between linguistic and visual

125   information. This approach reflects a somewhat passive construal of how people allocate

126   visual attention during language comprehension. In contrast, goal-based accounts of vision

127   start from the idea that eye movements reflect an active information-gathering process where

128   visual fixations are driven by task goals (Hayhoe & Ballard, 2005).

129   Under this account, people allocate visual attention to reduce uncertainty about the

130   world and maximize the expected future reward. For example, Hayhoe and Ballard (2005)

131   review evidence that people fixate on locations that are most helpful for their current goal

132   (an upcoming obstacle) as opposed to other aspects of a visual scene that might be more

133   salient (a flashing light). Moreover, other work shows that people gather task-specific

134   information via different visual routines as they become useful for their goals. For example,

135   Triesch et al 2003 found that people were much 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 the role of learning gaze patterns. 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 parts of the scene. For example, Shinoda, Hayhoe, and Shrivastava (2001) show that drivers, with practice, 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). Moreover, the timing of eye movements in these tasks often occur before an expected event (i.e., anticipatory), suggesting that gaze patterns reflect an interaction between people's expectations, information available in the visual scene, and their task goals.

Recent theoretical work has argued for a stronger link between goal-based perspectives and work on eye movements during language comprehension. For example, Salverda et al. (2011) highlight the immediate relevance of visual information with respect to the goal of language understanding, suggesting that listeners' goals should be a key predictor of fixation patterns. Moreover, they point out that factors such as the difficulty of executing a real world task should change decisions about where to look during language comprehension. One example of starting from a goal-based approach comes from Nelson and Cottrell (2007)' study of gaze patterns during category learning. Nelson and Cottrell (2007) modeled eye movements as a type of question-asking about features of a concept and showed that the dynamics of eye movements changed as participants became more familiar with the novel concepts. Early in learning people generated a broader distribution of fixations to explore all features. Later in learning, eye movements shifted to focus on a single stimulus dimension to maximize accuracy on the task. This shift in gaze patterns, from exploratory to efficient, suggests that fixation behavior changed as a function of changes in learning goals during the experiment.

163   In the current studies, the goal-based model of eye movements predicts that gaze

164 dynamics during language comprehension should adapt to the processing context. That is,

165 listeners should change the timing and location of eye movements when fixation locations

166 become more useful for language understanding. This proposal dovetails with a growing

167 body of research that explores the role of multisensory information available in face-to-face

168 communication such as gesture, prosody, facial expression and body movement.

169 **Language perception as multisensory integration**

170   Language comprehension is not just one stream of linguistic information. Instead,

171 face-to-face communication provides access to a set of multimodal cues that can facilitate

172 comprehension and there is a growing emphasis on studying language as a multimodal and

173 multisensory process (for a review, see Vigliocco, Perniss, and Vinson (2014)). For example,

174 empirical work shows that when gesture and speech provide redundant cues to meaning,

175 people are faster to process the information and make fewer errors (Kelly, Özyürek, & Maris,

176 2010). Moreover, developmental work shows that parents use visual cues such as gesture and

177 eye gaze to help structure language interactions with their children (Estigarribia & Clark,

178 2007). Finally, from a young age, children also produce gestures such as reaches and points

179 to share attention with others to achieve communicative goals (Liszkowski, Brown,

180 Callaghan, Takada, & De Vos, 2012).

181   Additional support for multisensory processing comes from work on audiovisual speech

182 perception, showing how spoken language perception is shaped by visual information coming

183 from a speaker's mouth. In a review, Peelle and Sommers (2015) point out that mouth

184 movements provide a clear indication of when someone has started to speak, which cues the

185 listener to allocate additional attention to the speech signal. Moreover, a speaker's mouth

186 movements convey information about the phonemes in the acoustic signal. For example,

187 visual speech information distinguishes between consonants such as /b/ vs. /d/ and place of

188 articulation can help a listener differentiate between words such as "cat" or "cap." Finally,

189   classic empirical work shows comprehension benefits for audiovisual speech compared to

190   auditory- or visual-only speech, especially in noisy listening contexts (Erber, 1969).

191   In sum, the work on multisensory processing shows that both auditory and visual

192   information interact to shape language perception. These results dovetail with the

193   interactive models of language processing reviewed earlier and suggest that visual

194   information can support comprehension (MacDonald & Seidenberg, 2006; McClelland et al.,

195   2006). Finally, these results highlight the value of studying language comprehension during

196   face-to-face communication, where listeners have the choice to gather visual information

197   about the linguistic signal from their social partners.

198   **The present studies**

199   The studies reported here synthesize ideas from research on language processing as a

200   multimodal, goal-based, and social phenomenon. We propose an information-seeking account

201   of eye movements during grounded language comprehension in social interaction. We

202   characterize the timing of gaze patterns as reflecting a tradeoff between gathering visual

203   information about the incoming linguistic signal from a speaker and seeking information

204   about the surrounding visual scene. We draw on models of eye movements as active

205   decisions that gather information to achieve reliable interpretations of incoming language.

206   We test predictions of our account using three case studies: sign language, text processing,

207   and processing spoken language in noisy environments. These case studies represent a broad

208   sampling of contexts that share a key feature: The interaction between the listener and

209   context changes the value of fixating on the language source to gather visual information for

210   comprehension.

211   A secondary goal of this work was to test whether children and adults would show

212   similar patterns of adaptation of gaze patterns. Recent developmental work shows that, like

213   adults, preschoolers will flexibly adjust how they interpret ambiguous sentences (e.g., "I had

214   carrots and *bees* for dinner.") by integrating information about the reliability of the incoming

215 perceptual information with their expectations about the speaker (Yurovsky, Case, & Frank,

216 2017). While children's behavior paralleled adults, they relied more on top-down

217 expectations about the speaker perhaps because their perceptual representations were noisier.

218 These developmental differences provide insight into how children succeed in understanding

219 language despite having partial knowledge of word-object mappings.

220   The structure of the paper is as follows. First, we compare children and adult's eye

221 movements while processing signed vs. spoken language. Then, we present a comparison of

222 adults' eye movements while processing serially printed text vs. spoken language. Finally, we

223 compare children and adults' gaze patterns while they process speech in noisy vs. clear

224 auditory environments. The key behavioral prediction (see Table 1 for more detailed

225 predictions) is that both children and adults will adapt the timing of their eye movements to

226 facilitate better word recognition. We hypothesized that when a language source provides

227 higher value visual information, listeners would prioritize fixations to their social partner,

228 and, in turn, would be slower to shift gaze away, generating (a) more accurate responses and

229 (b) fewer random, exploratory eye movements to the objects in the visual scene.

230   Before describing the studies, it is worth motivating our analytic approach. To quantify

231 the evidence for our predictions, we analyze the accuracy and reaction times (RTs) of

232 listeners' initial gaze shifts after hearing the name of an object. The timescale of this

233 analysis is milliseconds and focuses on a single decision within a series of decisions about

234 where to look during language processing. We chose this approach because first shifts are

235 rapid decisions driven by accumulating information about the identity of the named object.

236 Moreover, these measurements provide a window onto changes in the underlying dynamics of

237 how listeners integrate linguistic and visual information to make fixation decisions. Finally,

238 by focusing our analysis on a specific decision, we could leverage models of decision making

239 developed over the past decades to quantify changes in the underlying dynamics of eye

240 movements in different processing contexts (see the analysis plan for more details).

₂₄₁   **Experiment 1**

₂₄₂   Experiment 1 provides an initial test of our adaptive tradeoffs account. We compared

₂₄₃ eye movements of children learning American Sign Language to children learning a spoken

₂₄₄ language using parallel real-time language comprehension tasks where children processed

₂₄₅ familiar sentences (e.g., "Where's the ball?") while looking at a simplified visual world with

₂₄₆ 3 fixation targets (a center stimulus that varied by condition, a target picture, and a

₂₄₇ distracter picture; see Fig 1). The spoken language data are a reanalysis of three

₂₄₈ unpublished data sets, and the ASL data are reported in MacDonald et al. (2018). We

₂₄₉ predicted that, compared to spoken language processing, processing ASL would increase the

₂₅₀ value of fixating on the language source and decrease the value of generating exploratory,

₂₅₁ nonlanguage-driven shifts even after the disambiguating point in the linguistic signal.

₂₅₂ **Analysis plan**

₂₅₃   First, we present behavioral analyses of First Shift Accuracy and Reaction Time (RT).

₂₅₄ RT corresponds to the latency to shift away from the central stimulus to either picture

₂₅₅ measured from the onset of the target noun. Accuracy corresponds to whether participants'

₂₅₆ first gaze shift landed on the target or the distracter picture. It is important to note that this

₂₅₇ analysis of accuracy does not measure the overall amount of time spent looking at the target

₂₅₈ vs. the distractor image – a measure typically used in analyses of the Visual World Paradigm.

₂₅₉ We chose to focus on first shifts to provide a window onto how processing contexts change

₂₆₀ the underlying dynamics of information gathering decisions. All analysis code can be found

₂₆₁ in the online repository for this project: https://github.com/kemacdonald/speed-acc.

₂₆₂   We used the `rstanarm` (Gabry & Goodrich, 2016) package to fit Bayesian

₂₆₃ mixed-effects regression models. The mixed-effects approach allowed us to model the nested

₂₆₄ structure of our data – multiple trials for each participant and item, and a

₂₆₅ within-participants manipulation – by including random intercepts for each participant and

₂₆₆ item, and a random slope for each item and noise condition. We used Bayesian estimation to

267 quantify uncertainty in our point estimates, which we communicate using a 95% Highest

268 Density Interval (HDI). The HDI provides a range of credible values given the data and

269 model. Finally, to estimate age-related differences, we fit two types of models: (1) age group

270 (adults vs. children) as a categorical predictor and (2) age as a continuous predictor

271 (measured in days) within the child sample.

272   Next, we present the two model-based analyses – the EWMA and DDM. The goal of

273 these models is to move beyond a description of the data and map behavioral differences in

274 eye movements to underlying psychological variables. The EWMA method models changes in

275 random shifting behavior as a function of RT. For each RT, the model generates two values:

276 a "control statistic" (CS, which captures the running average accuracy of first shifts) and an

277 "upper control limit" (UCL, which captures the pre-defined limit of when accuracy would be

278 categorized as above chance level). Here, the CS is an expectation of random shifting to

279 either the target or the distracter image (nonlanguage-driven shifts), or a Bernoulli process

280 with probability of success 0.5. As RTs get slower, we assume that participants have

281 gathered more information and should become more accurate (language-driven), or a

282 Bernoulli process with probability success > 0.5. Using this model, we can quantify the

283 proportion of gaze shifts that were language-driven as opposed to random responding.

284   Following Vandekerckhove and Tuerlinckx (2007), we selected shifts categorized as

285 language-driven by the EWMA and fit a hierarchical Bayesian drift-diffusion model

286 (HDDM). The DDM quantifies differences in the underlying decision process that lead to

287 different patterns of behavior. The model assumes that people accumulate noisy evidence in

288 favor of one alternative with a response generated when the evidence crosses a pre-defined

289 decision threshold. We chose to implement a hierarchical Bayesian version of the DDM using

290 the HDDM Python package (Wiecki, Sofer, & Frank, 2013) since we had relatively few trials

291 from child participants and recent simulation studies have shown that the HDDM approach

292 was better than other DDM fitting methods for small data sets (Ratcliff & Childers, 2015).

293 Here, we focus on two parameters of interest: *boundary separation*, which indexes the

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

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).

Methods

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

   Spoken English samples. Participants were 80 native, monolingual English-learning children divided across three samples. Participants had no reported history of developmental or language 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
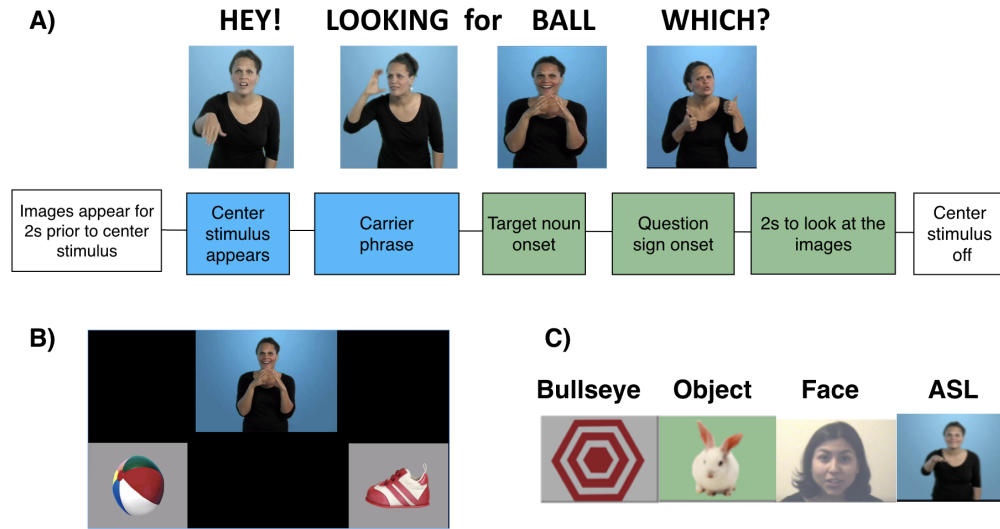
Figure 1. Stimuli for Experiment 1. Panel A shows the timecourse of the linguistic stimuli for a single trial. Panel B shows the layout of the fixation locations for all tasks: the center stimulus, the target, and the distracter. Panel C shows the four center stimulus items: a static geometric shape (Bullseye), a static image of a familiar object (Object), a person speaking (Face), and a person signing (ASL).

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 1.25 sec, ranging from 0.69 sec to 1.98 sec.

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

321 and said, "Look! Where's the (target word)?" Mean word length was 0.79 sec, ranging from

322 0.60 sec to 0.94 sec.

323   *ASL and English visual stimuli.* The image set consisted of colorful digitized pictures

324 of objects presented in fixed pairs with no phonological overlap (ASL task: cat—bird,

325 car—book, bear—doll, ball—shoe; English tasks: book-shoe, juice-banana, cookie-ball). Side

326 of target picture was counterbalanced across trials.

327   *Trial structure.* On each trial, the child saw two images of familiar objects on the

328 screen for two seconds before the center stimulus appeared. This time allowed the child to

329 visually explore both images. Next, the target sentence – which consisted of a carrier phrase,

330 target noun, and question sign – was presented, followed by two seconds without language to

331 allow the child to respond to the signer's sentence. The trial structure of the Face, Object,

332 and Bullseye tasks were highly similar: children were given two seconds to visually explore

333 the objects prior to the appearance of the center stimulus, then processed a target sentence,

334 and finally were given two seconds of silence to generate a response to the target noun.

335   **Design and procedure.** Children sat on their caregiver's lap and viewed the task

336 on a screen while their gaze was recorded using a digital camcorder. On each trial, children

337 saw two images of familiar objects on the screen for two seconds before the center stimulus

338 appeared (see Figure 1). Then they processed the target sentence – which consisted of a

339 carrier phrase, a target noun, and a question – followed by two seconds without language to

340 allow for a response. Participants saw 32 test trials with several filler trials interspersed to

341 maintain interest.

342   *Coding.* Participants' gaze patterns were coded (33-ms resolution) as being fixated on

343 either the center stimulus, one of the images, shifting between pictures, or away. To assess

344 inter-coder reliability, 25% of the videos were re-coded. Agreement was scored at the level of

345 individual frames of video and averaged 98% on these reliability assessments.
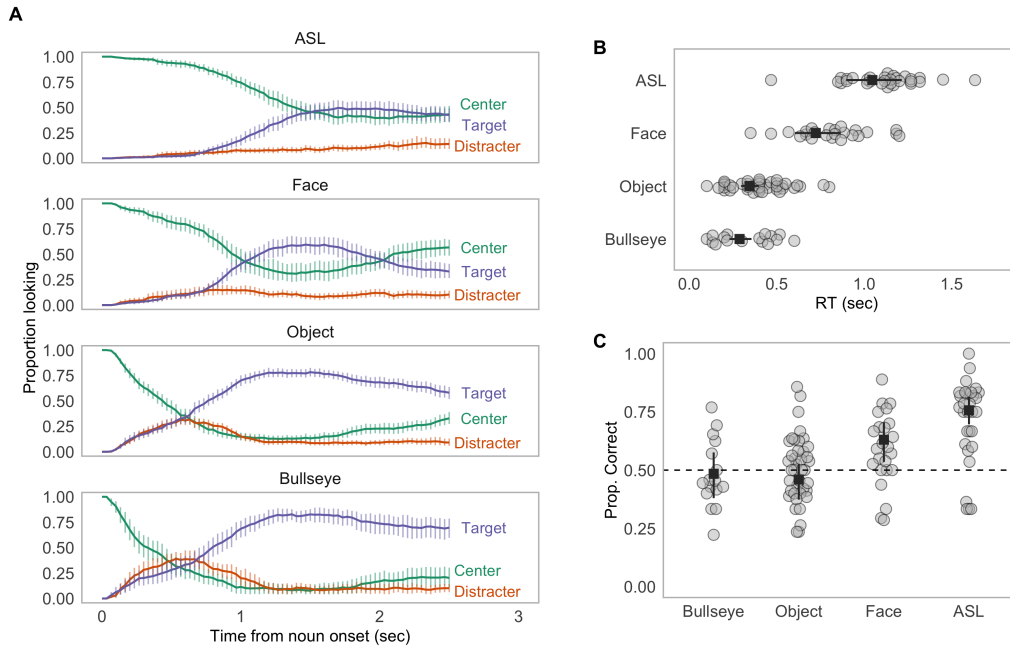
346 **Results and Discussion**

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. Color represents the processing context. Panel C shows the same information but for first shift accuracy.

Behavioral analyses. *Time course looking.* The first question of interest was how do young ASL and English learners distribute attention while processing language in real-time? Panel A of Fig. 2 presents an overview of children's looking to the center stimulus, target, and distractor images for each processing context. This plot shows changes in the mean proportion of trials on which participants fixated the signer, the target image, or the distracter image at every 33-ms interval of the stimulus sentence. At target-noun onset, children were looking at the center on all trials. 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 processing

357 contexts.

358   After looking to the target image, participants tended to shift their gaze back to the
359 center, shown by the increase in proportion looking to the center around two seconds after
360 target-noun onset. There were several qualitative differences in children's looking behavior
361 across the different center stimulus types. First, ASL-learners and English-learners who
362 processed sentences from a video of speaker spent more time looking to the center as
363 indicated by the shallower slope on their center-looking curves. Second, when the center
364 stimulus was a static geometric object (Bullseye) or a static familiar object (Object), spoken
365 language learners were more likely to look at the distractor image, especially early in the
366 time course of the target noun as indicated by the parallel increase in target and
367 distractor-looking curves in Fig. 2. In contrast, spoken language learners in the Face
368 context spent less time looking at the disracter, and ASL-learners rarely looked to the
369 distractor image at any point in the trial. This pattern of behavior provides qualitative
370 evidence that the children were adpating the dynamics of their looking depending on the
371 nature of the visual world and language modality.

372   *RT*. Panel B of Fig. 2 shows the full RT data distribution and the full posterior
373 distribution of the estimated RT difference between the noise and clear conditions. To
374 quantify differences across the groups, we fit a Bayesian linear mixed-effects regression
375 predicting first shift RT as a function of center stimulus type, controlling for age, and
376 including user-defined contrasts to test specific comparisons of interest: `Log(RT) ∼ center`
377 `stimulus type + age + (1 | subject) + (1 | item)`. We found that (a) ASL learners
378 generated slower RTs compared to all of the spoken English samples ($\beta = 0.60$ sec, 95% HDI
379 [0.44 sec, 0.76 sec]), (b) ASL learners' shifts were slower compared directly to children
380 processing spoken language in the Face condition ($\beta = 0.32$ sec, 95% HDI [0.13 sec, 0.52
381 sec]), and (c) children in the Face condition shifted gaze slower compared to participants in
382 the Object and Bullseye tasks ($\beta = 0.41$ sec, 95% HDI [0.29 sec, 0.55 sec]).

383   *Accuracy.* Next we compared the accuracy of first shifts across the different tasks by

384 fitting a mixed-effects logistic regression with the same specifications and contrasts as the RT

385 model. We found that (a) ASL learners were more accurate compared to all of the spoken

386 English samples ($\beta = 0.23$ sec, 95% HDI [0.17], 0.29), (b) ASL learners were more accurate

387 when directly compared to participants in the Face task ($\beta = 0.13$ sec, 95% HDI [0.04 sec,

388 0.23 sec]), (c) children learning spoken language were more accurate when processing

389 language from dynamic video of a person speaking compared to the Object and Bullseye

390 tasks ($\beta = 0.16$ sec, 95% HDI [0.07 sec, 0.24 sec]), and (d) English-learners' first shifts were

391 no different from random responding in the Object ($\beta = -0.04$ sec, 95% HDI [-0.13 sec, 0.03

392 sec]) and Bullseye ($\beta = -0.02$ sec, 95% HDI [-0.12 sec, 0.08 sec]) contexts.

393   **Model-based analyses.**   *EWMA.* Panel A of Fig. 3 shows changes in the control

394 statistic (CS) and the upper control limit (UCL) as a function of RT. Each CS starts at

395 chance performance and below the UCL. In the ASL and Face tasks, the CS value begins to

396 increase with RTs around 0.7 seconds after noun onset and eventually crosses the UCL,

397 indicating that responses > 0.7 sec were on average above chance levels. In contrast, the CS

398 in the Object and Bullseye tasks never crossed the UCL, indicating that children's shifts

399 were equally likely to land on the target or the distracter, regardless of when they were

400 initiated. This result suggests that first shifts in the Bullseye/Object tasks were not

401 language-driven and may instead have reflected a different process such as gathering more

402 information about the referents in the visual world.

403   Next, we compared the EWMA model fits for participants in the ASL and Face

404 processing contexts. We found that ASL learners generated fewer shifts when the CS was

405 below the UCL compared to children learning spoken language ($\beta = 0.14$, 95% HDI [0.08,

406 0.23]). This result indicates that ASL-learners were more likely to have gathered sufficient

407 information about the linguistic signal prior to shifting gaze away from the language source

408 (i.e., gaze shifts were language-driven). We found some evidence that ASL learners started

409 producing language-driven shifts earlier in the RT distribution as indicated by the point at

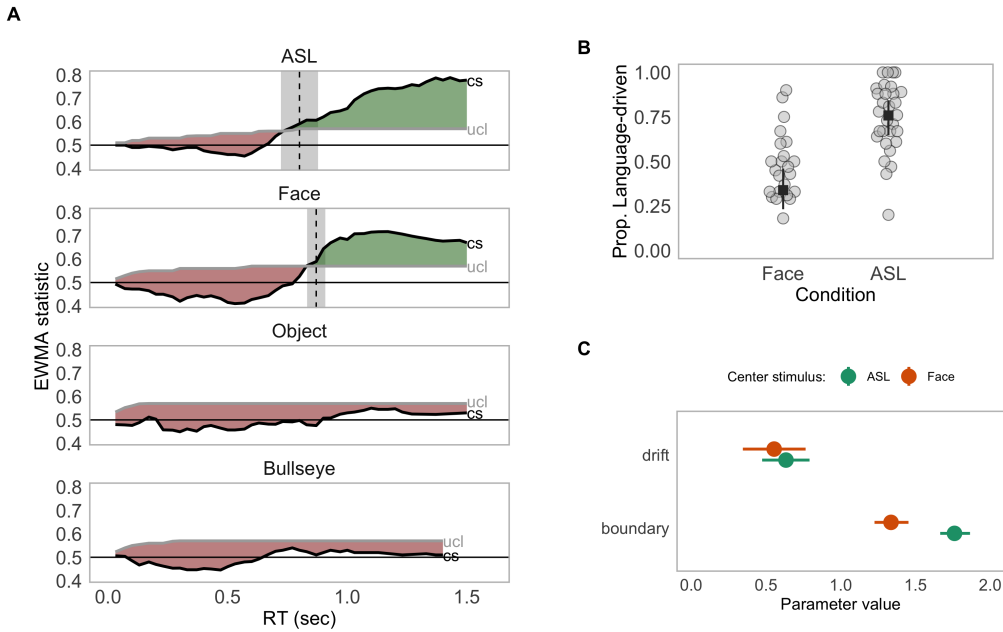410 which the CS crossed the UCL ($\beta = 0.22$ ms, 95% HDI [0.05 ms, 0.39 ms]).

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). The vertical dashed line is the median cutoff value (point when the control process shifts out of a guessing state). The grey shaded area represents the 95% confidence interval around the estimate of the median cutoff point, and the shaded ribbons represent the proportion of responses that were categorized as guesses (red) and language-driven (green). 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 posterior distributions for the boundary and drift rate parameters for the Face and ASL processing contexts.

₄₁₁   *HDDM.* Using the output of the EWMA, we compared the timing and accuracy of

₄₁₂ language-driven shifts for participants in the ASL and Face tasks.[1] We found that ASL

---

[1]We report the mean and the 95% highest density interval (HDI) of the posterior distributions for each parameter. The HDI represents the range of credible values given the model specification and the data. We chose not to interpret the DDM fits for the Bullseye/Face tasks since there was no suggestion of any

learners had a higher estimate for the boundary separation parameter compared to the Face

participants (ASL boundary = 1.76, HDI = [1.65, 1.88]; Face boundary = 1.34, HDI = [1.21,

1.47]), with no overlap in the credible values (see Fig 4). This suggests that ASL learners

accumulated 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, HDI = [0.44,

0.82]; Face drift = 0.55, HDI = [0.30, 0.80]).

Results summary. Taken together, the behavioral analyses and the EWMA/HDDM

results 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,

children learning ASL prioritized accuracy over speed, produced fewer nonlanguage-driven

shifts away from the center stimulus, and were more accurate with these gaze shifts.

Importantly, we did not see evidence in the HDDM model fits that these accuracy differences

could be explained by differential rates of information accumulation. Instead, the

model-based analyses suggest that ASL learners increased their decision threshold for

generating a response.

We hypothesized that this prioritization of gathering additional information is an

adaptive response to the channel competition present when processing a visual-manual

language. That is, when ASL learners shift gaze away a signer, they are deciding to leave an

area of the visual world that provides a great deal of useful and interesting information.

Moreover, unlike children leanring spoken languages, ASL learners cannot gather more of the

linguistic signal while looking at the objects. Thus, an adaptive language comprehension

system would increase levels of certainty before generating a response to maintain robust

understanding.

It is important to point out that these findings are based on exploratory analyses, and

our information seeking account was developed to explain this pattern of results. There are,

non-guessing signal.

439 however, several, potentially important differences between the stimuli, apparatus, and

440 populations that limit the sterngth of our interpretation of these data and the generality of

441 our account. Thus, in Experiments 2 and 3, we set out to perform well-controlleds,

442 confirmatory tests of our adaptive information seeking account of eye movements during

443 grounded language comprehension.

<p align="center">444   **Experiment 2**</p>

445   In Experiment 2, we aimed to replicate a key finding from Experiment 1: that

446 increasing the competition between fixating the language source and the nonlinguistic visual

447 world reduces nonlanguage-driven eye movements. Moreover, we conducted a confirmatory

448 test of our hypothesis that also controlled for the population differences present in

449 Experiment 1. We tested a sample of English-speaking adults using a within-participants

450 manipulation of the center stimulus type. We used the Face and Bullseye stimulus sets from

451 Experiment 1 and added two new conditions: Text, where the verbal language information

452 was accompanied by a word-by-word display of printed text (see Figure 3), and

453 Text-no-audio, where the spoken language stimulus was removed. We chose text processing

454 since, like sign language comprehension, information relevant to the linguistic signal is

455 concentrated in one location in the visual scene.

456   Our key behavioral prediction is that processing serially-presented text will shift the

457 value of allocating fixations to the center stimulus as the linguistic information unfolds in

458 time. This shift in information value should result in listeners allocating more fixations to

459 the center stimulus and fewer to the objects in the visual scene. This behavioral pattern

460 should be indexed by proportion guessing and cutoff point parameters of the EWMA model.

461 We did not have strong predictions for first shift accuracy and reaaction time or the DDM

462 parameter fits since the goal of the text manipulation was to modulate participants' strategic

463 allocation of visual attention and not the accuracy/efficiency of information processing.

464 **Methods**

465   **Participants.**   25 Stanford undergraduates participated (5 male) for course credit.

466 All participants were monolingual, native English speakers and had normal vision.

467   **Stimuli.**   Audio and visual stimuli were identical to the Face and Bullseye tasks in

468 Experiment 1. We included a new center fixation stimulus type: printed text. The text was

469 displayed in a white font on a black background and was programmed such that only a single

470 word appeared on the screen, with each word appearing for the same duration as the

471 corresponding word in the spoken language stimuli.

472   **Design and procedure.**   The design was nearly identical to Experiment 1, with the

473 exception of a change to a within-subjects manipulation where each participant completed

474 all four tasks (Bullseye, Face, Text, and Text-no-audio). In the Text condition, spoken

475 language accompanied the printed text. In the Text-no-audio condition, the spoken language

476 stimulus was removed. Participants saw a total of 128 trials while their eye movements were

477 tracked using automated eye-tracking software.

478 **Results and Discussion**

479   **Behavioral analyses.**   *Time course looking.* The first question of interest was how

480 do young ASL and English learners distribute attention while processing language in

481 real-time? Panel A of Fig. 4 presents an overview of adults's looking to the center stimulus,

482 target, and distractor images for each center stimulus type. Similar to children's looking

483 behavior in Experiment 1, at target-noun onset the majority of adults were looking at the

484 center. As the target noun unfolded, the mean proportion looking to the center decreased

485 rapidly as participants shifted their gaze to the objects. Proportion looking to the target

486 increased sooner and reached a higher asymptote compared to proportion looking to the

487 distracter for all four processing contexts.

488   After looking to the target image, adults did not tend to shif their gaze back to the

489 center as shown by the relatively flat proportion looking to the center curves one second

490 after target-noun onset. The primary qualitative difference in adult's looking behavior across
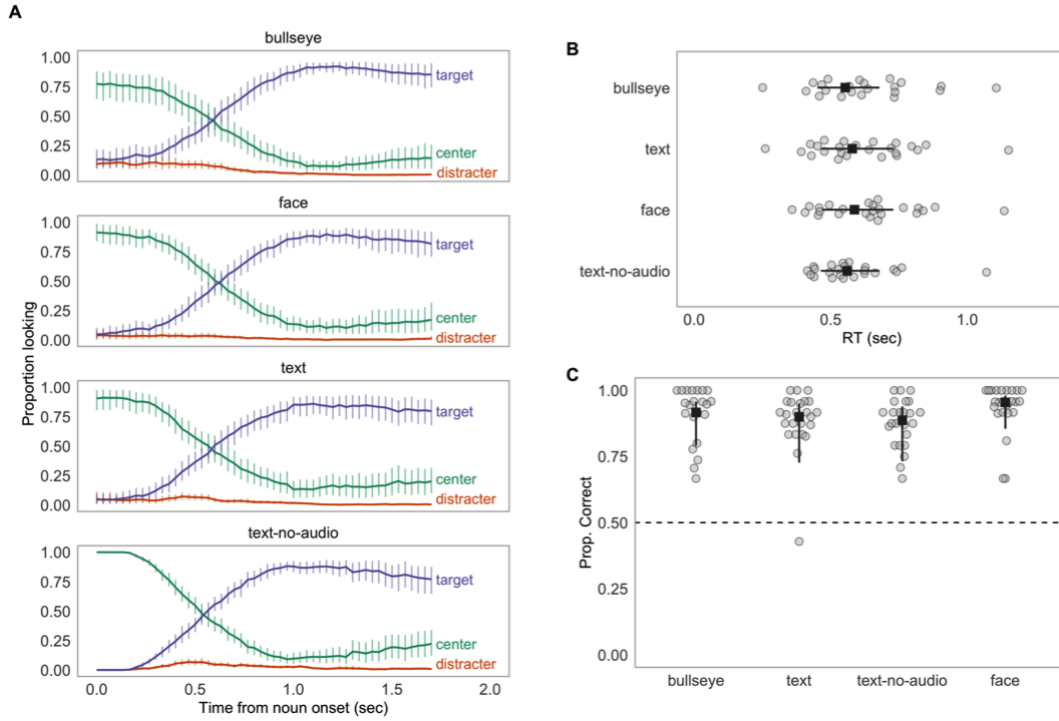
Figure 4. Results for the model-based analyses in Experiment 2. All plotting conventions are the same as in Figure 2.

the different center stimulus types was a higher tendency to be looking at the center stimulus in the Face and Text conditions relative to the Bullseye condition. This was especially true for the Text-no-audio condition where adults were looking to the center at target-noun onset on 100% of the trials. This pattern of behavior provides preliminary, qualitative evidence that the condition manipulation affected the dynamics of adults looking beahvior early in the time course of the target noun. We did not see evidence of differences in the asymptotes or slopes for the target- or distracter-looking curves.

   RT. Visual inspection of Figure 5, panel C suggests that mean response times of first shifts were similar across the four center stimulus conditions ($M_{bull} = 0.55$ sec, $M_{face} = 0.59$ sec, $M_{text} = 0.58$ sec, $M_{textNoaudio} = 0.56$ sec). We fit a linear mixed-effects regression with the same specification as in Experiment 1, but we added by-subject intercepts and slopes for each center stimulus type to account for our within-subjects manipulation. We did not see evidence that RTs were different across conditions, with the null value of zero condition

504 differences falling within the 95% CIs for each comparison of interest (see table XX in the

505 appendix for full model output).

506   *Accuracy.* Next, we modeled accuracy using a mixed-effects logistic regression with the

507 same specifications (see Panel B of Figur 5). We found that adults' first shifts were highly

508 accurate ($M_{bullseye} = 0.92$, $M_{face} = 0.95$, $M_{text} = 0.90$, $M_{textNoAudio} = 0.89$). And, in

509 contrast to the children in Experiment 1, adults' responses were above chance level even in

510 the Bullseye condition when the center stimulus was not salient or informative.

511   Adults' accurate first shifts suggests an interesting developmental difference in the

512 construal of the center stimulus in our task. This is speculative, but it seems plausible that

513 adults thought the Bullseye was designed to be a valid starting point for fixating gaze while

514 the sentence unfolded (i.e., someone put this here for a reason). As a result, if adults

515 maintained their fixation on the center stimulus for enough time to gather sufficient

516 linguistic singal, then they were highly accurate across all four processing condition, which is

517 reasonable since these were highly familiar words presented in child-directed speech.

518   Visual inspection of the timecourse looking curves, however, suggests that the effect of

519 the text manipulations occurred earlier in timecourse of decisions about visual fixation. That

520 is, in the first 300 ms after the start of the target word, adults in the Bullseye, Face, and

521 Text conditions, where they had access to linguistic information via the auditory channel,

522 were already allocating fixations away from the center stimulus and to the objects. In

523 contrast, in the Text-No-Audio condition, all of adults' fixation were directed to the center

524 stimulus location, which contained the language-relevant information. Next, we use our

525 model-based analyses to quantify these differences in adults' decisions about where to fixate

526 as a function of time.

527   **Model-based analyses.**   *EWMA.* For all four conditions, the control statistic

528 crossed the upper control limit (see Panel A of Figure 6), suggesting that at some point in

529 the RT distribution adults' shifts were reliably driven by linguistic information. Interestingly,

530 we found a graded effect of condition on the cutoff point (see the shift in the vertical dashed
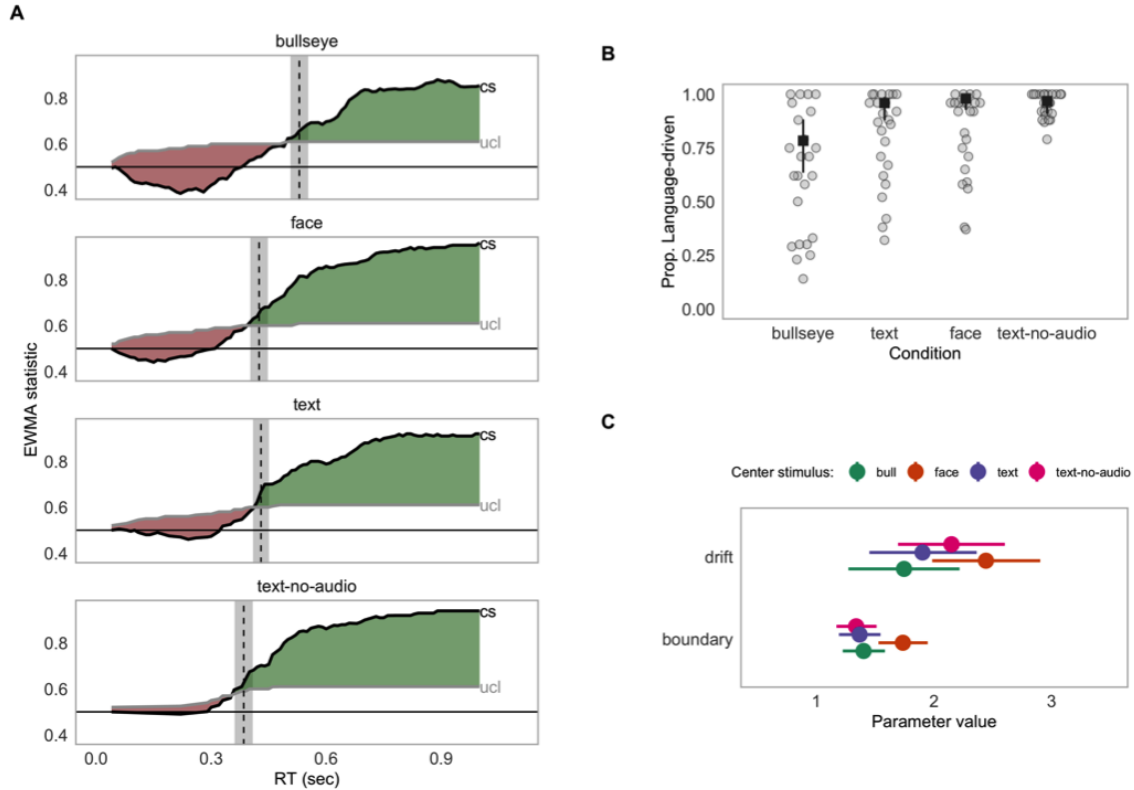
Figure 5. Results for the model-based analyses of Experiment 2. All plotting conventions are the same as Figure 3.

lines in Panel A of Figure 5). That is, the CS crossed the UCL earliest in the Text-no-audio condition ($M_{text-no-audio} = 0.39$, 95% HDI [0.37, 0.41]), followed by the Text ($M_{text} = 0.44$, 95% HDI [0.42, 0.46]) and Face ($M_{face} = 0.45$, 95% HDI [0.43, 0.47]) conditions, and finally the Bullseye condition ($M_{bullseye} = 0.54$, 95% HDI [0.52, 0.56]).[2]

We also found a smiliar pattern of a graded difference in the proportion of shifts that occurred when the control statistic was below the upper control limit ($M_{bullseye} = 0.78$, $M_{text} = 0.86$, $M_{text-no-audio} = 0.89$, $M_{face} = 0.93$). Adults generated fewer language-driven eye movements in the Bullseye condition comapred to the other contexts ($\beta = -0.12$, 95% HDI [-0.26, -0.01]). And the highest proportion of language-driven shifts in the Text-no-audio context ($\beta = 0.04$, 95% HDI [-0.02, 0.08]). These results provide evidence for our key prediction: that increasing the value of fixating the center stimulus for gathering linguistic

---

[2]See Table XX in the appendix for the relevant statistics for the pairwise comparisons of interest.

information reduced gaze shifts to the rest of the visual world. This shift in gaze dynamics, in turn, resulted in adults gathering more of the linguistic signal prior to generating eye movements away from the center stimulus, leading to a higher proporiton of language-driven shifts earlier in the distribution of reaction times.

HDDM. Using the classifications generated by the EWMA, we fit a HDDM to the language-drven shifts with the same specifications as in Experiment 1. There was high overlap of the posterior distributions for the drift rate parameters (see panel C of Figure 5), suggesting that participants gathered the linguistic information with similar efficiency. We also found high overlap in the distribution of boundary separation estimates for the Bullseye, Text, and Text-no-audio conditions. We saw some evidence for a higher boundary separation in the Face condition compared to the other three center stimulus types (Face boundary = 1.73, HDI = [1.49, 1.98]; Bullseye boundary = 1.40, HDI = [1.19, 1.62]; Text boundary = 1.37, HDI = [1.16, 1.58]; Text-no-audio boundary = 1.34, HDI = [1.14, 1.55]), indicating that adults' higher accuracy in this condition was driven by accumulating more information before generating a response. Note that the higher boundary separation and drift rate parameters for the Face condition differs from the results of the standard Accuracy analyses, which found similar patterns of performance. This occurs because the HDDM estimates parameter fits using reaction times distributions for both correct and incorrect responses.

Results summary. Together, these results suggest that adults were sensitive to the tradeoff between gathering different kinds of visual information. When processing text, people generated fewer nonlanguage-driven shifts (EWMA results) but their processing efficiency of the linguistic signal itself did not change (HDDM results). Interestingly, we found a graded difference in the EWMA results between the Text and Text-no-audio conditions, with the lowest proportion of early, nonlanguage-driven shifts occurring while processing text without the verbal stimuli. This behavior makes sense; if the adults could rely on the auditory channel to gather the linguistic information, then the value of fixating the text display decreases. In contrast to the children in Experiment 1, adults were highly

accurate in the Bullseye condition, perhaps because they construed the Bullseye as a center fixation that they *should* fixate, or perhaps they had better encoded the location/identity of the two referents prior to the start of the target sentence.

Experiment 3

In this experiment, we recorded 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 predicted that processing speech in a noisy context would make participants less likely to shift before collecting sufficient information.[3] This delay, in turn, would lead to a lower proportion of shifts flagged as random/exploratory in the EWMA analysis, and a pattern of DDM results indicating a prioritization of accuracy over and above speed of responding (see the Analysis Plan section below for more details on the models). We also 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 the same pattern of adapting gaze patterns to gather more visual information in the noisy processing context.

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).

---

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

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

600 *Noise manipulation.* To create the stimuli in the noise condition, we convolved each
601 recording with Brown noise using the Audacity audio editor. The average signal-to-noise
602 ratio[4] in the noise condition was 2.87 dB compared to the clear condition, which was 35.05
603 dB.

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

608 **Design and procedure.** Participants viewed the task on a screen while their gaze
609 was tracked using an SMI RED corneal-reflection eye-tracker mounted on an LCD monitor,
610 sampling at 60 Hz. The eye-tracker was first calibrated for each participant using a 6-point
611 calibration. On each trial, participants saw two images of familiar objects on the screen for
612 two seconds before the center stimulus appeared (see Fig. **??**). Next, they processed the
613 target sentence – which consisted of a carrier phrase, a target noun, and a question –
614 followed by two seconds without language to allow for a response. Child participants saw 32
615 trials (16 noise trials; 16 clear trials) with several filler trials interspersed to maintain
616 interest. Adult participants saw 64 trials (32 noise; 32 clear). The noise manipulation was
617 presented in a blocked design with the order of block counterbalanced across participants.

618 **Results and discussion**

---

[4]The ratio of signal power to the noise power, with values greater than 0 dB indicating more signal than
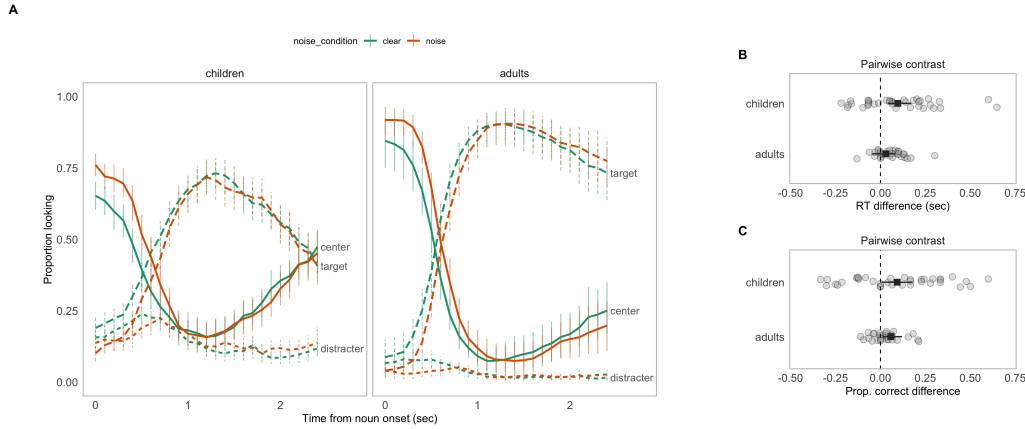noise.

Figure 6. Behavioral results for children and adults in Experiment 3. 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.

₆₁₉   **Behavioral analyses:.   RT.** To make RTs more suitable for modeling on a linear

₆₂₀ scale, we analyzed responses in log space with the final model specified as:

₆₂₁ $log(RT) \sim noise\_condition + age\_group + (noise\_condition \mid sub\_id) + (noise\_condition \mid$

₆₂₂ $target\_item)$. Panel A of Figure **??** shows the full RT data distribution and the full

₆₂₃ posterior distribution of the estimated RT difference between the noise and clear conditions.

₆₂₄ Both children and adults were slower to identify the target in the noise condition (Children

₆₂₅ $M_{noise} = 500.19$ sec; Adult $M_{noise} = 595.23$ sec), as compared to the clear condition

₆₂₆ (Children $M_{clear} = 455.72$ sec Adult $M_{clear} = 542.45$ sec). RTs in the noise condition were

₆₂₇ 48.82 seconds slower on average, with a 95% HDI ranging from 3.72 sec to 96.26 ms, and not

₆₂₈ including the null value of zero condition difference. Older children responded faster than

₆₂₉ younger children (M_{age} = -0.44, [-0.74, -0.16]), with little evidence for an interaction between
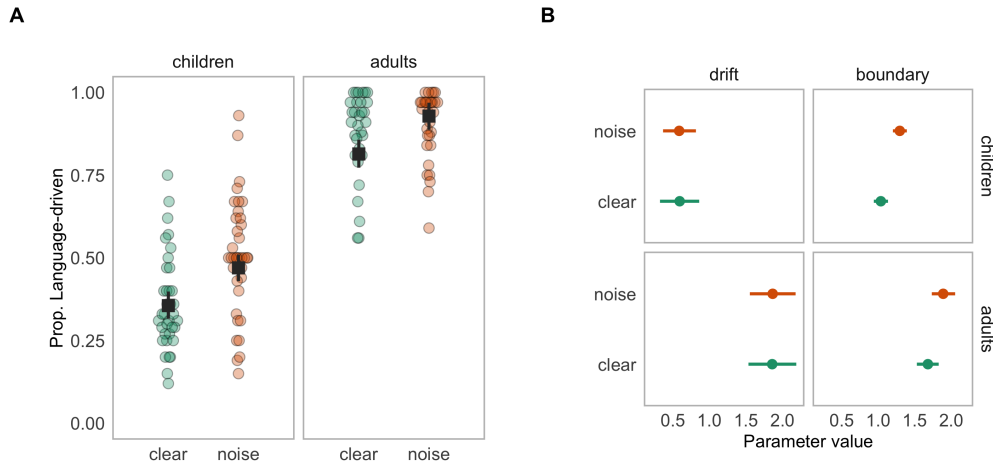
₆₃₀ age and condition.

Figure 7. Results for the model-based analyses for Experiment 3. The majority of plotting conventions are the same as Figure 3. In Panel C, linetype and alpha value represent age group: children vs. adults.

⁶³¹ **Accuracy.** Next, we modeled adults and children's first shift accuracy using a

⁶³² mixed-effects logistic regression with the same specifications (see Panel B of Fig. **??**). Both

⁶³³ groups were more accurate than a model of random responding (null value of 0.5 falling well

⁶³⁴ outside the lower bound of the 95% HDI for all group means). Adults were more accurate

⁶³⁵ ($M_{adults} = 90\%$) than children ($M_{children} = 61\%$). The key result is that both groups showed

⁶³⁶ evidence of higher accuracy in the noise condition: children ($M_{noise} = 67\%$; $M_{clear} = 61\%$)

⁶³⁷ and adults ($M_{noise} = 92\%$; $M_{clear} = 90\%$). Accuracy in the noise condition was on average

⁶³⁸ 4% higher, with a 95% HDI from -1% to 12%. Note that the null value of zero difference falls

⁶³⁹ at the very edge of the HDI. But 95% of the credible values are greater than zero, providing

⁶⁴⁰ evidence for higher accuracy in the noise condition. Within the child sample, there was no

⁶⁴¹ evidence of a main effect of age or an interaction between age and noise condition.

⁶⁴² **Model-based analyses:. EWMA.** Fig. **??** shows the proportion of shifts that the

⁶⁴³ model classified as random vs. language-driven for each age group and processing context.

⁶⁴⁴ On average, 41% (95% HDI: 32%, 50%) of children's shifts were categorized as

⁶⁴⁵ language-driven, which was significantly fewer than adults, 87% (95% HDI: 78%, 96%).

646 Critically, processing speech in a noisy context caused both adults and children to generate a

647 higher proportion of language-driven shifts (i.e., fewer random, exploratory shifts away from

648 the speaker), with the 95% HDI excluding the null value of zero condition difference ($\beta_{noise} =$

649 11%, [7.00%, 16%]). Within the child sample, older children generated fewer random, early

650 shifts ($M_{age} = $ -0.21, [-0.35, -0.08]). There was no eivdence of an interaction between age and

651 condition. This pattern of results suggests that the noise condition caused participants to

652 increase visual fixations to the language source, leading them to generate fewer exploratory,

653 random shifts before accumulating sufficient information to respond accurately.

654 **HDDM.** Fig. **??** shows the full posterior distributions for the HDDM output.

655 Children had lower drift rates (children $M_{drift} = $ NA; adults $M_{drift} = $ NA) and boundary

656 separation estimates (children $M_{boundary} = 1.02$; adults $M_{boundary} = 1.33$) as compared to

657 adults, suggesting that children were less efficient and less cautious in their responding. The

658 noise manipulation selectively affected the boundary separation parameter, with higher

659 estimates in the noise condition for both age groups ($\beta_{noise} = 0.26$, [0.10, 0.42]). This result

660 suggests that participants' in the noise condition prioritized information accumulation over

661 speed when generating an eye movement in response to the incoming language. This

662 increased decision threshold led to higher accuracy. Moreover, the high overlap in estimates

663 of drift rate suggests that participants were able to integrate the visual and auditory signals

664 such that they could achieve a level of processing efficiency comparable to the clear

665 processing context.

666 Taken together, the behavioral and EWMA/HDDM results provide converging support

667 for the predictions of our information-seeking account. Processing speech in noise caused

668 listeners to seek additional visual information to support language comprehension. Moreover,

669 we observed a very similar pattern of behavior in children and adults, with both groups

670 producing more language-driven shifts and prioritizing accuracy over speed in the more

671 challenging, noisy environment.

General Discussion

672

673   Language comprehension in grounded, social contexts involves extracting meaning from

674   the linguistic signal and mapping it to the surrounding world. But how should listeners

675   prioritize integrating different kinds of visual information? In this work, we proposed that

676   listeners flexibly adapt their gaze behaviors in response to features of the social context,

677   seeking visual information from their social partners when it was especially useful for

678   language comprehension. We present evidence for our account using three case studies

679   sampled from a diverse set of language processing contexts: sign language, dynamic displays

680   of printed text, and spoken language in noisy auditory environments. We found that,

681   compared to children learning spoken English, young ASL-learners delayed their gaze shifts

682   away from a language source, were more accurate with these shifts, and produced a smaller

683   proportion of random shifting behavior. Next, English-speaking adults produced fewer

684   random gaze shifts when processing dynamic displays of printed text compared to spoken

685   language. Finally, 3-5 year-olds and adults delayed the timing of gaze shifts away from a

686   speaker's face when processing speech in a noisy environment, which resulted in fewer

687   random eye movements and more accurate gaze shifts. Together, these results provide

688   evidence that young listeners, like adults, adapt their gaze patterns to the demands of

689   different processing environments by seeking out visual information from social partners to

690   support language comprehension.

691   This work attempts to integrate top-down, goal-based models of vision (Hayhoe &

692   Ballard, 2005) with work on language-driven eye movements (Allopenna et al., 1998). While

693   we chose to start with two case studies – ASL and text processing – we think the account is

694   more general and that there are many real world situations where people must negotiate the

695   tradeoff between gathering more information about language or about the world: e.g.,

696   processing spoken language in noisy environments or at a distance; or early in language

697   learning when children are acquiring new words and often rely on nonlinguistic cues to

698   reference such as pointing or eye gaze. Overall, we hope this work contributes to a broader

account of eye movements during language comprehension that can explain fixation

behaviors across a wider variety of populations, processing contexts, and during different

stages of language learning.

These results synthesize ideas from several research programs, including work on

language-mediated visual attention (Tanenhaus et al., 1995), goal-based accounts of vision

during everyday tasks (Hayhoe & Ballard, 2005), and work on effortful listening (Van Engen

& Peelle, 2014). Moreover, our findings parallel recent work by McMurray, Farris-Trimble,

and Rigler (2017) showing 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 to gaze patterns in adults with typical hearing by degrading the auditory

stimuli so that it shared features with the output of a cochlear implant (noise-vocoded

speech).

The results reported here also dovetail with recent developmental work by Yurovsky et

al. (2017). In that 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. Here, 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 more accurate responses in the more

challenging, noisy context.

Limitations

This work has several important limitations that pave the way for future work. 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 wider 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.

Conclusion

This experiment tested the generalizability of our information-seeking account of eye movements within the domain of grounded language comprehension. But the account 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, this work integrates goal-based models of eye-movements with language comprehension in grounded, social contexts. This approach presents a way forward for explaining fixation behaviors across a wider variety processing contexts and during different stages of language learning.

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.

Byers-Heinlein, K., Morin-Lessard, E., & Lew-Williams, C. (2017). Bilingual infants control their languages as they listen. *Proceedings of the National Academy of Sciences, 114*(34), 9032–9037.

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.

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.

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.

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

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

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

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).

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.

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.

Schwartz, R. G., Steinman, S., Ying, E., Mystal, E. Y., & Houston, D. M. (2013). Language processing in children with cochlear implants: A preliminary report on lexical access for production and comprehension. *Clinical Linguistics & Phonetics*, *27*(4), 264–277.

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.

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

Van Engen, K. J., & Peelle, J. E. (2014). Listening effort and accented speech. *Frontiers in*

829   *Human Neuroscience, 8.*

830   Venker, C. E., Eernisse, E. R., Saffran, J. R., & Weismer, S. E. (2013). Individual differences

831   in the real-time comprehension of children with asd. *Autism Research, 6*(5), 417–432.

832   Vigliocco, G., Perniss, P., & Vinson, D. (2014). Language as a multimodal phenomenon:

833   Implications for language learning, processing and evolution. The Royal Society.

834   Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical bayesian estimation of

835   the drift-diffusion model in python. *Frontiers in Neuroinformatics, 7*, 14.

836   Yee, E., & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic

837   activation during spoken word recognition. *Journal of Experimental Psychology:*

838   *Learning, Memory, and Cognition, 32*(1), 1.

839   Yurovsky, D., Case, S., & Frank, M. C. (2017). Preschoolers flexibly adapt to linguistic input

840   in a noisy channel. *Psychological Science, 28*(1), 132–140.