

# The development of children's ability to track and predict turn structure in conversation

Marisa Casillas<sup>a,\*</sup>, Michael C. Frank<sup>b</sup>

<sup>a</sup>*Max Planck Institute for Psycholinguistics, Nijmegen*

<sup>b</sup>*Department of Psychology, Stanford University*

---

## Abstract

Children begin developing turn-taking skills in infancy but take several years to assimilate their growing knowledge of language into their turn-taking behavior. In two eye-tracking experiments, we measured children's anticipatory gaze to upcoming responders while controlling linguistic cues to turn structure. In Experiment 1, we showed English and non-English conversations to English-speaking adults and children. In Experiment 2, we phonetically controlled lexicosyntactic and prosodic cues in English-only speech. Children spontaneously made anticipatory gaze switches by age two and continued improving through age six. In both experiments, children and adults made more anticipatory switches after hearing questions. Like adults, prosody alone did not improve children's predictive gaze shifts. But, unlike adults, lexical cues alone were not sufficient to improve prediction—children's performance was

---

\*Corresponding author

best overall with lexicosyntax and prosody together. Our findings support an account in which turn prediction emerges in infancy, but then only gradually becomes fully integrated with linguistic processing.

*Keywords:* Turn taking, Conversation, Development, Prosody, Questions, Eye-tracking

---

<sup>1</sup> **1. Introduction**

<sup>2</sup> Spontaneous conversation is a universal context for using and learning  
<sup>3</sup> language. Like other types of human interaction, it is organized at its core  
<sup>4</sup> by the roles and goals of its participants. But, what sets conversation apart is  
<sup>5</sup> its structure: sequences of interconnected, communicative actions that take  
<sup>6</sup> place across alternating turns at talk. Sequential, turn-based structures in  
<sup>7</sup> conversation are strikingly uniform across language communities and linguis-  
<sup>8</sup> tic modalities. Turn-taking behaviors are also cross-culturally consistent in  
<sup>9</sup> their basic features and the details of their implementation (De Vos et al.,  
<sup>10</sup> 2015; Dingemanse et al., 2013; Stivers et al., 2009).

<sup>11</sup> Children participate in sequential coordination (proto-turn taking) with  
<sup>12</sup> their caregivers starting at three months of age—before they can rely on  
<sup>13</sup> any linguistic cues (see, among others, Bateson, 1975; Hilbrink et al., 2015;  
<sup>14</sup> Jaffe et al., 2001; Snow, 1977). However, infant turn taking is different from

15 adult turn taking in several ways: it is heavily scaffolded by caregivers, has  
16 different timing from adult turn taking, and lacks semantic content (Hilbrink  
17 et al., 2015; Jaffe et al., 2001). But children’s early, turn-structured social  
18 interactions are presumably a critical precursor to their later conversational  
19 turn taking. Early non-verbal interactions likely establish the protocol by  
20 which children come to use language with others. How do children integrate  
21 linguistic knowledge with these preverbal turn-taking abilities, and how does  
22 this integration change over the course of childhood?

23 In this study, we investigate when children begin to make predictions  
24 about upcoming turn structure in conversation, and how they integrate lan-  
25 guage into their predictions as they grow older. In the remainder of the  
26 introduction, we first give a basic review of turn-taking research and the  
27 state of current knowledge about adult turn prediction. We then discuss  
28 recent work on the development of turn-taking skills before turning to the  
29 details of our own study.

30 *1.1. Adult turn taking*

31 Turn taking itself is not unique to conversation. Many other human activ-  
32 ities are organized around sequential turns at action. Traffic intersections and  
33 computer network communication both use turn-taking systems. Children’s

<sup>34</sup> early games (e.g., give-and-take, peek-a-boo) have built-in, predictable turn  
<sup>35</sup> structure (Ratner and Bruner, 1978; Ross and Lollis, 1987). Even monkeys  
<sup>36</sup> take turns: non-human primates such as marmosets and Campbell's monkeys  
<sup>37</sup> vocalize contingently with each other in both natural and lab-controlled en-  
<sup>38</sup> vironments (Lemasson et al., 2011; Takahashi et al., 2013). In all these cases,  
<sup>39</sup> turn taking serves as a protocol for interaction, allowing the participants to  
<sup>40</sup> coordinate with each other through sequences of contingent action.

<sup>41</sup> Conversational turn taking distinguishes itself from other turn-taking be-  
<sup>42</sup> haviors by the complexity of the sequencing involved. Conversational turns  
<sup>43</sup> come grouped into semantically-contingent sequences of action. The groups  
<sup>44</sup> can span turn-by-turn exchanges (e.g., simple question-response, "How are  
<sup>45</sup> you?"—"Fine.") or sequence-by-sequence exchanges (e.g., reciprocals, "How  
<sup>46</sup> are you?"—"Fine, and you?"—"Great!"). Compared to other turn-taking be-  
<sup>47</sup> haviors, the possible sequence and action types in everyday talk are can be  
<sup>48</sup> diverse and unpredictable.

<sup>49</sup> Despite this complexity, conversational turn taking is precise in its timing.  
<sup>50</sup> Across a diverse sample of conversations in 10 languages, one study found  
<sup>51</sup> a consistent average turn transition time of 0–200 msec at points of speaker  
<sup>52</sup> switch (Stivers et al., 2009). Experimental results and current models of

53 speech production suggest that it takes approximately 600 msec to produce  
54 a content word, and even longer to produce a simple utterance (Griffin and  
55 Bock, 2000; Levelt, 1989). So in order to achieve 200 msec turn transitions,  
56 speakers must begin formulating their response before the prior turn has  
57 ended (Levinson, 2013, 2016). Moreover, to formulate their response early  
58 on, speakers must track and anticipate what types of response might become  
59 relevant next. They also need to predict the content and form of upcoming  
60 speech so that they can launch their articulation at exactly the right moment.  
61 Prediction thus plays a key role in timely turn taking.

62 Adults have a lot of information at their disposal to help make accurate  
63 predictions about upcoming turn content. Lexical, syntactic, and prosodic  
64 information (e.g., *wh*- words, subject-auxiliary inversion, and list intonation)  
65 can all inform addressees about upcoming linguistic structure (De Ruiter  
66 et al., 2006; Duncan, 1972; Ford and Thompson, 1996; Torreira et al., 2015).  
67 Non-verbal cues (e.g., gaze, posture, and pointing) often appear at turn-  
68 boundaries and can sometimes act as late indicators of an upcoming speaker  
69 switch (Rossano et al., 2009; Stivers and Rossano, 2010). Additionally, the  
70 sequential context of a turn can make it clear what will come next: answers  
71 after questions, thanks or denial after compliments, etc. (Schegloff, 2007).

72 Prior work suggests that adult listeners primarily use lexicosyntactic in-  
73 formation to accurately predict upcoming turn structure. De Ruiter and  
74 colleagues (2006) asked participants to listen to snippets of spontaneous con-  
75 versation and to press a button whenever they anticipated that the current  
76 speaker was about to finish his or her turn. The speech snippets were con-  
77 trolled for the amount of linguistic information present; some were normal,  
78 but others had flattened pitch, low-pass filtered speech, or further manip-  
79 ulations. With pitch-flattened speech, the timing of participants' button  
80 responses was comparable to their timing with the full linguistic signal. But  
81 when no lexical information was available, participants' responses were sig-  
82 nificantly earlier. The authors concluded that lexicosyntactic information<sup>1</sup>  
83 was necessary and possibly sufficient for turn-end projection, while intona-  
84 tion was neither necessary nor sufficient. Congruent evidence comes from  
85 studies varying the predictability of lexicosyntactic and pragmatic content:  
86 adults anticipate turn ends better when they can more accurately predict the  
87 exact words that will come next (Magyari and De Ruiter, 2012; see also Mag-

---

<sup>1</sup>The “lexicosyntactic” condition only included flattened pitch and so was not exclusively lexicosyntactic—the speech would still have residual prosodic structure, including syllable duration and intensity.

<sup>88</sup> yari et al., 2014). They can also identify speech acts within the first word of  
<sup>89</sup> an utterance (Gísladóttir et al., 2015), allowing them to start planning their  
<sup>90</sup> response at the first moment possible (Bögels et al., 2015).

<sup>91</sup> Despite this body of evidence, the role of prosody for adult turn predic-  
<sup>92</sup> tion is still a matter of debate. De Ruiter and colleagues' (2006) experiment  
<sup>93</sup> focused on the role of intonation, which is only a partial index of prosody.  
<sup>94</sup> Prosody is tied closely to the syntax of an utterance, so the two linguistic  
<sup>95</sup> signals are difficult to control independently (Ford and Thompson, 1996).  
<sup>96</sup> Torreira, Bögels and Levinson (2015) used a combination of button-press  
<sup>97</sup> and verbal responses to investigate the relationship between lexicosyntac-  
<sup>98</sup> tic and prosodic cues in turn-end prediction. Critically, their stimuli were  
<sup>99</sup> cross-spliced so that each item had full prosodic cues to accompany the lex-  
<sup>100</sup> icosyntax. Because of the splicing, they were able to create items that had  
<sup>101</sup> syntactically-complete units with no intonational phrase boundary at the  
<sup>102</sup> end. Participants never verbally responded or pressed the “turn-end” but-  
<sup>103</sup> ton when hearing a syntactically-complete phrase without an intonational  
<sup>104</sup> phrase boundary. And when intonational phrase boundaries were embedded  
<sup>105</sup> in multi-utterance turns, participants were tricked into pressing the “turn-  
<sup>106</sup> end” button 29% of the time. Their results suggest that listeners actually

<sup>107</sup> do rely on prosodic cues to execute a response (see also de De Ruiter et al.  
<sup>108</sup> (2006):525). These experimental findings corroborate other corpus and ex-  
<sup>109</sup> perimental work promoting a combination of cues (lexicosyntactic, prosodic,  
<sup>110</sup> and pragmatic) as key for accurate turn-end prediction (Duncan, 1972; Ford  
<sup>111</sup> and Thompson, 1996; Hirvenkari et al., 2013).

<sup>112</sup> *1.2. Turn taking in development*

<sup>113</sup> The majority of work on children's early turn taking has focused on ob-  
<sup>114</sup> servations of spontaneous interaction. Children's first turn-like structures  
<sup>115</sup> appear as early as two to three months after birth, in proto-conversation with  
<sup>116</sup> their caregivers (Bruner, 1975, 1985). During proto-conversations, caregivers  
<sup>117</sup> treat their infants as capable of making meaningful contributions: they take  
<sup>118</sup> every look, vocalization, arm flail, and burp as "utterances" in the joint dis-  
<sup>119</sup> course (Bateson, 1975; Jaffe et al., 2001; Snow, 1977). Infants catch onto the  
<sup>120</sup> structure of proto-conversations quickly. By three to four months they notice  
<sup>121</sup> disturbances to the contingency of their caregivers' response and, in reaction,  
<sup>122</sup> change the rate and quality of their vocalizations (Bloom, 1988; Masataka,  
<sup>123</sup> 1993; Toda and Fogel, 1993).

<sup>124</sup> The timing of children's responses to their caregivers' speech shows a  
<sup>125</sup> non-linear pattern. Infants' contingent vocalizations in the first few months

126 of life show very fast timing (though with a lot of vocal overlap). But by  
127 nine months, their timing slows down considerably, only to gradually speed  
128 up again after 12 months (Hilbrink et al., 2015). For children, taking turns  
129 with brief transitions between speakers is more difficult than avoiding speaker  
130 overlap; children's incidence of overlap is nearly adult-like by nine months,  
131 but the timing of their non-overlapped responses stays much longer than  
132 the adult 200 msec standard for the next few years (Casillas et al., In press;  
133 Garvey, 1984; Garvey and Berninger, 1981; Ervin-Tripp, 1979). This puzzling  
134 pattern is likely due to their linguistic development: taking turns on time  
135 is easier when the response is a simple vocalization rather than a linguistic  
136 utterance. Integrating language into the turn-taking system may therefore  
137 be one major factor in children's delayed responses (Casillas et al., In press).

138 Children, like adults, might use linguistic cues in the ongoing turn to  
139 make predictions about upcoming turn structure. Studies of early linguistic  
140 development point to a possible early advantage for prosody over lexicosyn-  
141 tax in children's turn-taking predictions. Infants can distinguish their native  
142 language's rhythm type from others soon after birth (Mehler et al., 1988;  
143 Nazzi and Ramus, 2003); they show preference for the typical stress patterns  
144 of their native language over others by 6–9 months (e.g., iambic vs. trochaic),

145 and can use prosodic information to segment the speech stream into smaller  
146 chunks from 8 months onward (Johnson and Jusczyk, 2001; Morgan and  
147 Saffran, 1995). Four- to five-month-olds also prefer pauses in speech to be  
148 inserted at prosodic boundaries, and by 6 months infants can use prosodic  
149 markers to pick out sub-clausal syntactic units, both of which are useful for  
150 extracting turn structure from ongoing speech (Jusczyk et al., 1995; Soder-  
151 strom et al., 2003). In comparison, children show at best a very limited  
152 lexical inventory before their first birthday (Bergelson and Swingley, 2013;  
153 Shi and Melancon, 2010).

154 Keitel and colleagues (2013) were one of the first to explore how children  
155 use linguistic cues to predict upcoming turn structure. They asked 6-, 12-,  
156 24-, and 36-month-old infants, and adult participants to watch short videos  
157 of conversation and tracked their eye movements at points of speaker change.  
158 They showed their participants two types of conversation videos—one nor-  
159 mal and one with flattened pitch (i.e., with flattened intonation contours)—  
160 to test the role of intonation in participants' anticipatory predictions about  
161 upcoming speech. Comparing children's anticipatory gaze frequency to a  
162 random baseline, they found that only 36-month-olds and adults made an-  
163 ticipatory gaze switches more often than expected by chance. Among those,

<sub>164</sub> only 36-month-olds were affected by a lack of intonation contours, leading  
<sub>165</sub> Keitel and colleagues to conclude that children’s ability to predict upcoming  
<sub>166</sub> turn structure relies on their ability to comprehend the stimuli lexicoseman-  
<sub>167</sub> tically. They also suggest that intonation might play a secondary role in turn  
<sub>168</sub> prediction, but only after children acquire more sophisticated, adult-like lan-  
<sub>169</sub> guage comprehension abilities (also see Keitel and Daum, 2015).

<sub>170</sub> Although the Keitel et al. (2013) study constitutes a substantial ad-  
<sub>171</sub> vance over previous work in this domain, it has some limitations. Because  
<sub>172</sub> these limitations directly inform our own study design, we review them in  
<sub>173</sub> some detail. First, their estimates of baseline gaze frequency (“random” in  
<sub>174</sub> their terminology) were not random. Instead, they used gaze switches dur-  
<sub>175</sub> ing ongoing speech as a baseline. But ongoing speech is the period in which  
<sub>176</sub> switching is least likely to occur (Hirvenkari et al., 2013)—their baseline thus  
<sub>177</sub> maximizes the chance of finding a difference in gaze frequency at turn transi-  
<sub>178</sub> tions compared to the baseline. A more conservative baseline would compare  
<sub>179</sub> participants’ looking behavior at turn transitions to their looking behavior  
<sub>180</sub> during randomly selected windows of time throughout the stimulus, includ-  
<sub>181</sub> ing turn transitions. We follow this conservative approach in the current  
<sub>182</sub> study.

183        Second, the conversation stimuli Keitel et al. (2013) used were some-  
184        what unusual. The average gap between turns was 900 msec, a duration  
185        much longer than typical adult timing, which averages around 200 msec  
186        (Stivers et al., 2009). The speakers in the videos were also asked to mini-  
187        mize their movements while performing scripted, adult-directed conversation,  
188        which would have created a somewhat unnatural interaction. Additionally,  
189        to produce more naturalistic conversation, it would have been ideal to lo-  
190        calize the sound sources for the two voices in the video (i.e., to have the  
191        voices come out of separate left and right speakers). But both voices were  
192        recorded and played back on the same audio channel, which may have made  
193        it difficult to distinguish the two talkers. Again, we attempt to address these  
194        issues in our current study. Despite these minor methodological issues, the  
195        Keitel et al. (2013) study still demonstrates intriguing age-based differences  
196        in children's ability to predict upcoming turn structure. Our current work  
197        takes this paradigm as a starting point.<sup>2</sup>

---

<sup>2</sup>But also see Casillas and Frank (2012, 2013).

198 *1.3. The current study*

199 Our goal in the current study is to find out when children begin to make  
200 predictions about upcoming turn structure and to understand how their pre-  
201 dictions are affected by linguistic cues across development. We present two  
202 experiments in which we measured children's anticipatory gaze to respon-  
203 ders while they watched conversation videos with natural (people speaking  
204 English vs. non-English; Experiment 1) and non-natural (puppets with pho-  
205 netically manipulated speech; Experiment 2) control over the presence of  
206 lexical and prosodic cues. We tested children across a wide range of ages  
207 (Experiment 1: 3–5 years; Experiment 2: 1–6 years), with adult control  
208 participants in each experiment.

209 We highlight three primary findings: first, although children and adults  
210 use linguistic cues to make predictions about upcoming turn structure, they  
211 do so primarily in predict speaker transitions after questions (a speech act  
212 effect). This intriguing effect, which has not been reported previously, sug-  
213 gests that participants track unfolding speech for cues to upcoming speaker  
214 change, which may affect how they use linguistic cues more generally for  
215 anticipatory processing in conversation. Second, we find that children make  
216 more predictions than expected by chance starting at age two, but that this

<sub>217</sub> effect is small at first, and continues to improve through age six. Third, we  
<sub>218</sub> find no evidence of an early prosody advantage in children's anticipations  
<sub>219</sub> and, further, no evidence that prosodic or lexical cues alone can substitute  
<sub>220</sub> for their combination in the full linguistic signal (as is proposed for adults;  
<sub>221</sub> De Ruiter et al., 2006). Instead, anticipation is strongest for stimuli with the  
<sub>222</sub> full range of cues. In sum, our findings support an account in which turn  
<sub>223</sub> prediction emerges in infancy, but becomes fully integrated with linguistic  
<sub>224</sub> processing only gradually across development.

<sub>225</sub> **2. Experiment 1**

<sub>226</sub> We recorded participants' eye movements as they watched six short videos  
<sub>227</sub> of two-person (dyadic) conversation interspersed with attention-getting filler  
<sub>228</sub> videos. Each conversation video featured an improvised discourse in one  
<sub>229</sub> of five languages (English, German, Hebrew, Japanese, and Korean). Par-  
<sub>230</sub> ticipants saw two videos in English and one in every other language. The  
<sub>231</sub> participants, all native English speakers, were only expected to understand  
<sub>232</sub> the two videos in English. We showed participants non-English videos to  
<sub>233</sub> limit their access to lexical information while maintaining their access to  
<sub>234</sub> other cues to turn boundaries (e.g., (non-native) prosody, gaze, inbreaths,

<sup>235</sup> phrase final lengthening). Using this method, we compared children and  
<sup>236</sup> adult's anticipatory looks from the current speaker to the upcoming speaker  
<sup>237</sup> at points of turn transition in English and non-English videos.

<sup>238</sup> *2.1. Methods*

<sup>239</sup> *2.1.1. Participants*

<sup>240</sup> We recruited 74 children between ages 3;0–5;11 and 11 undergraduate  
<sup>241</sup> adults to participate in the experiment. Our child sample included 19 three-  
<sup>242</sup> year-olds, 32 four-year-olds, and 23 five-year-olds, all enrolled in a local nurs-  
<sup>243</sup> ery school. All participants were native English speakers. Approximately  
<sup>244</sup> one-third (N=25) of the children's parents and teachers reported that their  
<sup>245</sup> child regularly heard a second (and sometimes third or further) language, but  
<sup>246</sup> only one child frequently heard a language that was used in our non-English  
<sup>247</sup> video stimuli, and we excluded his data from analyses. None of the adult  
<sup>248</sup> participants reported fluency in a second language.

<sup>249</sup> *2.1.2. Materials*

<sup>250</sup> *Video recordings.* We recorded pairs of talkers while they conversed in  
<sup>251</sup> a sound-attenuated booth (see a sample frame in Figure 1). Each talker  
<sup>252</sup> was a native speaker of the language being recorded, and each talker pair



Figure 1: Example frame from a conversation video used in Experiment 1.

253 was male-female. Using a Marantz PMD 660 solid state field recorder, we  
254 captured audio from two lapel microphones, one attached to each participant,  
255 while simultaneously recording video from the built-in camera of a MacBook  
256 laptop computer. The talkers were volunteers and were acquainted with their  
257 recording partner ahead of time.

258 Each recording session began with a 20-minute warm-up period of spon-  
259 taneous conversation during which the pair talked for five minutes on four  
260 topics (favorite foods, entertainment, hometown layout, and pets). Then we  
261 asked talkers to choose a new topic—one relevant to young children (e.g.,  
262 riding a bike, eating breakfast)—and to improvise a dialogue on that topic.  
263 We asked them to speak as if they were on a children’s television show in  
264 order to elicit child-directed speech toward each other. We recorded until the

265 talkers achieved at least 30 seconds of uninterrupted discourse with enthu-  
266 siastic, child-directed speech. Most talker pairs took less than five minutes  
267 to complete the task, usually by agreeing on a rough script at the start. We  
268 encouraged talkers to ask at least a few questions to each other during the  
269 improvisation. The resulting conversations were therefore not entirely spon-  
270 taneous, but were as close as possible while still remaining child-oriented in  
271 topic, prosodic pattern, and lexicosyntactic construction.<sup>3</sup>

272 After recording, we combined the audio and video recordings by hand,  
273 and cropped each recording to the (approximate) 30-second interval with the  
274 most turn activity. Because we recorded the conversations in stereo, the male  
275 and female voices came out of separate speakers during video playback. This  
276 gave each voice in the videos a localized source (from the left or right loud-  
277 speaker). We coded each turn transition in the videos for language condition  
278 (English vs. non-English), inter-turn gap duration (in milliseconds), and  
279 speech act (question vs. non-question). The non-English stimuli were coded  
280 for speech act from a monolingual English-speaker's perspective, i.e., which

---

<sup>3</sup>All of the non-English talkers were fluent in English as a second language, and some fluently spoke three or more languages. We chose male-female pairs as a natural way of creating contrast between the two talker voices.

281 turns “sound like” questions, and which do not: we asked five native Amer-  
282 ican English speakers to listen to the audio recording for each non-English  
283 turn and judge whether it sounded like a question. We marked non-English  
284 turns as questions when at least 4 of the 5 listeners (80%) said that the turn  
285 “sounded like a question”.

286 Because the conversational stimuli were recorded semi-spontaneously, the  
287 duration of turn transitions and the number of speaker transitions in each  
288 video was variable. We measured the duration of each turn transition from  
289 the audio recording associated with each video. We excluded turn transi-  
290 tions longer than 550 msec and shorter than 90 msec, also excluding over-  
291 lapped transitions, from analysis.<sup>4</sup> This left approximately equal numbers  
292 of turn transitions available for analysis in the English (N=20) and non-  
293 English (N=16) videos. On average, the inter-turn gaps for English videos

---

<sup>4</sup>Overlap occurs when a responder begins a new turn before the current turn is finished. When overlap occurs, observers cannot switch their gaze in anticipation of the response because the response began earlier than expected. Participants expect conversations to proceed with “one speaker at a time” (Sacks et al., 1974). They would therefore still be fixated on the prior speaker when the overlap started, and would have to switch their gaze *reactively* to the responder.

294 (mean=318, median=302, stdev=112 msec) were slightly longer than for non-  
295 English videos (mean=286, median=251, stdev=122 msec). The longer gaps  
296 in the English videos could give them a slight advantage: our definition of  
297 an “anticipatory gaze shift” includes shifts that are initiated during the gap  
298 between turns (Figure 2), so participants had slightly more time to make  
299 anticipatory shifts in the English videos.

300 Questions made up exactly half of the turn transitions in the English  
301 (N=10) and non-English (N=8) videos. In the English videos, inter-turn  
302 gaps were slightly shorter for questions (mean=310, median=293, stdev=112  
303 msec) than non-questions (mean=325, median=315, stdev=118 msec). Non-  
304 English videos did not show a large difference in transition time for questions  
305 (mean=270, median=257, stdev=116 msec) and non-questions (mean=302,  
306 median=252, stdev=134 msec).

307 *2.1.3. Procedure*

308 Participants sat in front of an SMI 120Hz corneal reflection eye-tracker  
309 mounted beneath a large flatscreen display. The display and eye-tracker were  
310 secured to a table with an ergonomic arm that allowed the experimenter to  
311 position the whole apparatus at a comfortable height, approximately 60 cm  
312 from the viewer. We placed stereo speakers on the table, to the left and right

313 of the display.

314 Before the experiment started, we warned adult participants that they  
315 would see videos in several languages and that, though they weren't expected  
316 to understand the content of non-English videos, we *would* ask them to an-  
317 swer general, non-language-based questions about the conversations. Then  
318 after each video we asked participants one of the following randomly-assigned  
319 questions: "Which speaker talked more?", "Which speaker asked the most  
320 questions?", "Which speaker seemed more friendly?", and "Did the speak-  
321 ers' level of enthusiasm shift during the conversation?" We also asked if the  
322 participants could understand any of what was said after each video. The  
323 participants responded verbally while an experimenter noted their responses.

324 Children were less inclined to simply sit and watch videos of conversation  
325 in languages they didn't speak, so we used a different procedure to keep them  
326 engaged: the experimenter started each session by asking the child about  
327 what languages he or she could speak, and about what other languages he  
328 or she had heard of. Then the experimenter expressed her own enthusiasm  
329 for learning about new languages, and invited the child to watch a video  
330 about "new and different languages" together. If the child agreed to watch,  
331 the experimenter and the child sat together in front of the display, with

332 the child centered in front of the tracker and the experimenter off to the  
333 side. Each conversation video was preceded and followed by a 15–30 second  
334 attention-getting filler video (e.g., running puppies, singing muppets, flying  
335 bugs). If the child began to look bored, the experimenter would talk during  
336 the fillers, either commenting on the previous conversation (“That was a neat  
337 language!”) or giving the language name for the next conversation (“This  
338 next one is called Hebrew. Let’s see what it’s like.”) The experimenter’s  
339 comments reinforced the video-watching as a joint task.

340 All participants (child and adult) completed a five-point calibration rou-  
341 tine before the first video started. We used a dancing Elmo for the chil-  
342 dren’s calibration image. During the experiment, participants watched all  
343 six 30-second conversation videos. The first and last conversations were in  
344 American English and the intervening conversations were Hebrew, Japanese,  
345 German, and Korean. The presentation order of the non-English videos was  
346 shuffled into four lists, which participants were assigned to randomly. The  
347 entire experiment, including instructions, took 10–15 minutes.

348 *2.1.4. Data preparation and coding*

349 To determine whether participants predicted upcoming turn transitions,  
350 we needed to define a set of criteria for what counted as an anticipatory gaze

shift. Prior work using similar experimental procedures has found that adults and children make anticipatory gaze shifts to upcoming talkers within a wide time frame; the earliest shifts occur before the end of the prior turn, and the latest occur after the onset of the response turn, with most shifts occurring in the inter-turn gap (Keitel et al., 2013; Hirvenkari, 2013; Tice and Henetz, 2011). Following prior work, we measured how often our participants shifted their gaze from the prior to the upcoming speaker *before* the shift in gaze could have been initiated in reaction to the onset of the speaker's response. In doing so, we assumed that it takes participants 200 msec to plan an eye movement, following standards from adult anticipatory processing studies (e.g., Kamide et al., 2003).

We checked each participant's gaze at each turn transition for three characteristics (Figure 2): (1) that the participant fixated on the prior speaker for at least 100 msec at the end of the prior turn, (2) that sometime thereafter the participant switched to fixate on the upcoming speaker for at least 100 ms, and (3) that the switch in gaze was initiated within the first 200 msec of the response turn, or earlier. These criteria guarantee that we only counted gaze shifts when: (1) participants were tracking the previous speaker, (2) switched their gaze to track the upcoming speaker, and (3) did so before

370 they could have simply reacted to the onset of speech in the response. Under  
371 this assumption, a gaze shift that was initiated within the first 200 msec of  
372 the response (or earlier) was planned *before* the child could react to the onset  
373 of speech itself.

374 As mentioned, most anticipatory switches happen in the inter-turn gap,  
375 but we also allowed anticipatory gaze switches that occurred in the final  
376 syllables of the prior turn. Early switches are consistent with the distribu-  
377 tion of responses in explicit turn-boundary prediction tasks. For example, in  
378 a button press task, adult participants anticipate turn ends approximately  
379 200 msec in advance of the turn's end, and anticipatory responses to pitch-  
380 flattened stimuli come even earlier (De Ruiter et al., 2006). We therefore  
381 allowed switches to occur as early as 200 msec before the end of the prior  
382 turn. Again, because it takes 200 msec to plan an eye movement, we counted  
383 anticipatory switches, at the latest, 200 msec after the onset of speech. There-  
384 fore, for very early and very late switches, our requirement of 100 msec of  
385 fixation on each speaker would sometimes extend outside of the transition  
386 window boundaries (200 msec before and after the inter-turn gap). The max-  
387 imally available fixation window was therefore 100 msec before and after the  
388 earliest and latest possible switch point (300 msec before and after the inter-

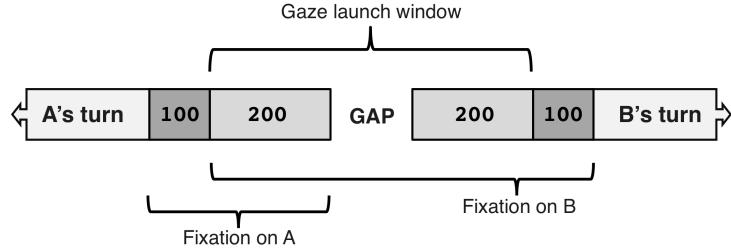


Figure 2: Schematic summary of criteria for anticipatory gaze shifts from speaker A to speaker B during a turn transition.

389 turn gap). We did not count switches made during the fixation window as  
 390 anticipatory. We *did* count switches made during the inter-turn gap. The  
 391 period of time from the beginning of the possible fixation window on the  
 392 prior speaker to the end of the possible fixation window on the responder  
 393 was our total analysis window (300 msec + the inter-turn gap + 300 msec).

394 *Predictions.* We expected participants to show greater anticipation in the  
 395 English videos than in the non-English videos because of their increased  
 396 access to linguistic information in English. We also predicted that anticipa-  
 397 tion would be greater following questions compared to non-questions; ques-  
 398 tions have early cues to upcoming turn transition (e.g., *wh*- words, subject-  
 399 auxiliary inversion), and also make a next response immediately relevant.  
 400 Our third prediction was that anticipatory looks would increase with devel-

<sup>401</sup> opment, along with children's increased linguistic competence.

<sup>402</sup> *2.2. Results*

<sup>403</sup> Participants looked at the screen most of the time during video playback  
<sup>404</sup> (81% and 91% on average for children and adults, respectively). They pri-  
<sup>405</sup> marily kept their eyes on the person who was currently speaking in both  
<sup>406</sup> English and non-English videos: they gazed at the current speaker between  
<sup>407</sup> 38% and 63% of the time, looking back at the addressee between 15% and  
<sup>408</sup> 20% of the time (Table 1). Even three-year-olds looked more at the current  
<sup>409</sup> speaker than anything else, whether the videos were in a language they could  
<sup>410</sup> understand or not. Children looked at the current speaker less than adults  
<sup>411</sup> did during the non-English videos. Despite this, their looks to the addressee  
<sup>412</sup> did not increase substantially in the non-English videos, indicating that their  
<sup>413</sup> looks away were probably related to boredom rather than confusion about  
<sup>414</sup> ongoing turn structure. Overall, participants' pattern of gaze to current  
<sup>415</sup> speakers demonstrated that they performed basic turn tracking during the  
<sup>416</sup> videos, regardless of language. Figure 3 shows participants' anticipatory gaze  
<sup>417</sup> rates across age, language condition, and transition type.

Age group	Condition	Speaker	Addressee	Other onscreen	Offscreen
3	English	0.61	0.16	0.14	0.08
4	English	0.60	0.15	0.11	0.13
5	English	0.57	0.15	0.16	0.12
Adult	English	0.63	0.16	0.16	0.05
3	Non-English	0.38	0.17	0.20	0.25
4	Non-English	0.43	0.19	0.21	0.18
5	Non-English	0.40	0.16	0.26	0.18
Adult	Non-English	0.58	0.20	0.16	0.07

Table 1: Average proportion of gaze to the current speaker and addressee during periods of talk.

<sup>418</sup> *2.2.1. Statistical models*

<sup>419</sup> We identified anticipatory gaze switches for all 36 usable turn transitions,  
<sup>420</sup> based on the criteria outlined in Section 2.1.4, and analyzed them for effects  
<sup>421</sup> of language, transition type, and age with two mixed-effects logistic regres-  
<sup>422</sup> sions (Bates et al., 2014; R Core Team, 2014). We built one model each  
<sup>423</sup> for children and adults. We modeled children and adults separately because  
<sup>424</sup> effects of age are only pertinent to the children’s data. The child model  
<sup>425</sup> included condition (English vs. non-English)<sup>5</sup>, transition type (question vs.

---

<sup>5</sup>Because each non-English language was represented by a single stimulus, we cannot treat individual languages as factors. Gaze behavior might be best for non-native languages that have the most structural overlap with participants’ native language: English speakers can make predictions about the strength of upcoming Swedish prosodic boundaries nearly

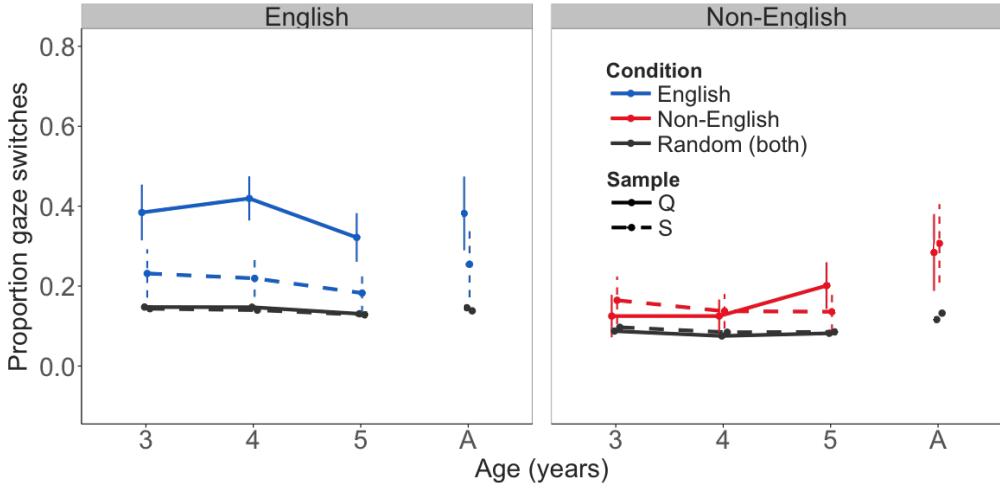


Figure 3: Anticipatory gaze rates across language condition and transition type for the real data (red and blue) and the randomly permuted baselines (gray). Vertical bars represent 95% confidence intervals.

<sup>426</sup> non-question), age (3, 4, 5; numeric), and duration of the inter-turn gap  
<sup>427</sup> (seconds, e.g., 0.441) as predictors, with full interactions between condition,  
<sup>428</sup> transition type, and age. We included the duration of the inter-turn gap as a  
<sup>429</sup> control predictor since longer gaps provide more opportunities to make antici-  
<sup>430</sup> patory switches (Figure 2). We additionally included random effects of item

---

as well as Swedish speakers do, but Chinese speakers are at a disadvantage in the same task (Carlson et al., 2005). We would need multiple items from each of the languages to check for similarity effects of specific linguistic features.

**Children**

	Estimate	Std. Error	<i>z</i> value	Pr(>  <i>z</i>  )
(Intercept)	-0.96145	0.84915	-1.132	0.257531
Age	-0.18268	0.17509	-1.043	0.296764
LgCond= <i>non-English</i>	-3.29349	0.96055	-3.429	0.000606 ***
Type= <i>non-Question</i>	-1.10131	0.86520	-1.273	0.203055
GapDuration	3.40171	1.22878	2.768	0.005634 **
Age*LgCond= <i>non-English</i>	0.52066	0.21192	2.457	0.014015 *
Age*Type= <i>non-Question</i>	-0.01628	0.19442	-0.084	0.933262
LgCond= <i>non-English</i> *	2.68171	1.35045	1.986	0.047057 *
Type= <i>non-Question</i>				
Age*LgCond= <i>non-English</i> *	-0.45633	0.30168	-1.513	0.130378
Type= <i>non-Question</i>				

**Adults**

	Estimate	Std. Error	<i>z</i> value	Pr(>  <i>z</i>  )
(Intercept)	-0.1966	0.6945	-0.283	0.777062
LgCond= <i>non-English</i>	-0.8812	0.9607	-0.917	0.359028
Type= <i>non-Question</i>	-4.4953	1.3147	-3.419	0.000628 ***
GapDuration	-1.1227	1.9889	-0.565	0.572414
LgCond= <i>non-English</i> *	3.2972	1.6115	2.046	0.040747 *
Type= <i>non-Question</i>				
LgCond= <i>non-English</i> *	1.3625	3.0097	0.453	0.650749
GapDuration				
Type= <i>non-Question</i> *	10.5107	3.3482	3.139	0.001694 **
GapDuration				
LgCond= <i>non-English</i> *	-6.3156	4.4969	-1.404	0.160191
Type= <i>non-Question</i> *				
GapDuration				

Table 2: Model output for children and adults' anticipatory gaze switches in Experiment 1.

431 (turn transition) and participant, with random slopes of condition, transition

<sup>432</sup> type, and their interaction for participants (Barr et al., 2013).<sup>6</sup> The adult  
<sup>433</sup> model included condition, transition type, duration, and their interactions as  
<sup>434</sup> predictors with participant and item included as random effects and random  
<sup>435</sup> slopes of condition, transition type, and their interaction for participant.

<sup>436</sup> Children's anticipatory gaze switches showed effects of language condition  
<sup>437</sup> ( $\beta=-3.29$ ,  $SE=0.961$ ,  $z=-3.43$ ,  $p<.001$ ) and gap duration ( $\beta=3.4$ ,  $SE=1.229$ ,  
<sup>438</sup>  $z=2.77$ ,  $p<.01$ ) with additional effects of an age-by-language condition in-  
<sup>439</sup> teraction ( $\beta=0.52$ ,  $SE=0.212$ ,  $z=2.46$ ,  $p<.05$ ) and a language condition-by-  
<sup>440</sup> transition type interaction ( $\beta=2.68$ ,  $SE=1.35$ ,  $z=1.99$ ,  $p<.05$ ). There were  
<sup>441</sup> no significant effects of age or transition type alone ( $\beta=-0.18$ ,  $SE=0.175$ ,  
<sup>442</sup>  $z=-1.04$ ,  $p=.3$  and  $\beta=-1.10$ ,  $SE=0.865$ ,  $z=-1.27$ ,  $p=.2$ , respectively).

<sup>443</sup> Adults' anticipatory gaze switches showed an effect of transition type  
<sup>444</sup> ( $\beta=-4.5$ ,  $SE=1.315$ ,  $z=-3.42$ ,  $p<.001$ ) and significant interactions between  
<sup>445</sup> language condition and transition type ( $\beta=3.3$ ,  $SE=1.61$ ,  $z=2.05$ ,  $p<.05$ )  
<sup>446</sup> and transition type and gap duration ( $\beta=10.51$ ,  $SE=3.348$ ,  $z=3.139$ ,  $p<.01$ ).

---

<sup>6</sup>The models we report are all qualitatively unchanged by the exclusion of their random slopes. We have left the random slopes in because of minor participant-level variation in the predictors modeled.

447 *2.2.2. Random baseline comparison*

448 We estimated the probability that these patterns were the result of ran-  
449 dom looking by running the same regression models on participants' real  
450 eye-tracking data, only this time calculating their anticipatory gaze switches  
451 with respect to randomly permuted turn transition windows. This process  
452 involved: (1) randomizing the order and temporal placement of the analysis  
453 windows within each stimulus (see Figure 4; “analysis window” is defined  
454 in Figure 2) to randomly redistribute the analysis windows across the eye-  
455 tracking signal, (2) re-running each participant's eye tracking data through  
456 switch identification (described in Section 2.1.4) on each of the randomly per-  
457 muted analysis windows, and (3) modeling the anticipatory switches from the  
458 randomly permuted data with the same statistical models we used for the  
459 original data (Section 2.2.1; Table 2). Importantly, although the onset time  
460 of each transition was shuffled within the eye-tracking signal, the other intrin-  
461 sic properties of each turn transition (e.g., prior speaker identity, transition  
462 type, gap duration, language condition, etc.) stayed constant across each  
463 random permutation.

464 This procedure effectively de-links participants' gaze data from the turn  
465 structure in the original stimulus, thereby allowing us to compare turn-

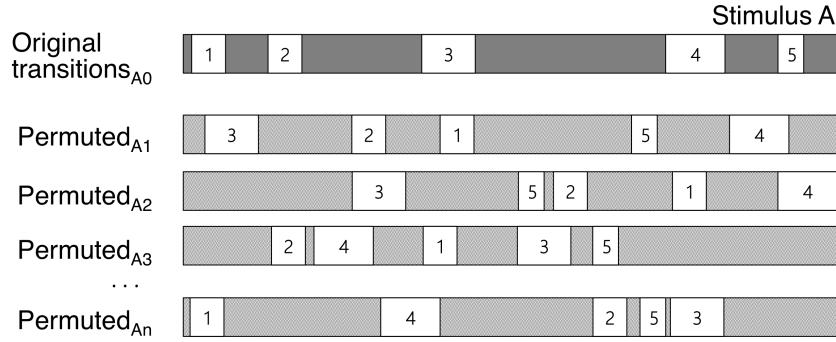


Figure 4: Example of analysis window permutations for a stimulus with five turn transitions. The windows were  $\pm 300$  msec around the inter-turn gap.

<sup>466</sup> related (original) and non-turn-related (randomly permuted) looking behav-  
<sup>467</sup> ior using the same eye movement data. The resulting anticipatory gazes from  
<sup>468</sup> the randomly permuted analysis windows represent an average anticipatory  
<sup>469</sup> gaze rate over all possible starting points: a random baseline.

<sup>470</sup> By running the real and randomly permuted data sets through identical  
<sup>471</sup> statistical models, we can estimate how likely it is that predictor effects in  
<sup>472</sup> the original data (e.g., the effect of language condition; Table 2) arose from  
<sup>473</sup> random looking. Because these analyses are complex, we report their full  
<sup>474</sup> details in Appendix A.

<sup>475</sup> Our baseline analyses revealed that none of the significant predictors  
<sup>476</sup> from models of the original, turn-related data can be explained by random

<sup>477</sup> looking. For the children's data, the original  $z$ -values for language condition, gap duration, the age-language condition interaction, and the language condition-transition type interaction were all greater than 95% of  $z$ -values for the randomly permuted data (99.9%, 95.5%, 99.4%, and 96%, respectively, all  $p < .05$ ). Similarly, the adults' data showed significant differentiation from the randomly permuted data for two of the three originally significant predictors—transition type and the transition type-gap duration interaction (greater than 99.9% and 99.7% of random  $z$ -values, respectively, all  $p < .01$ )—with marginal differentiation for the interaction of language condition and transition type (greater than 94.6% of random  $z$ -values;  $p = .054$ ).

<sup>487</sup> See Section Appendix A for more information on each predictor's random permutation distribution.

<sup>489</sup> *2.2.3. Developmental effects*

<sup>490</sup> The models reported above revealed a significant interaction of age and language condition (Table 2) that was unlikely be due to random looking (Figure 3). To further explore this effect, we compared the effect of language condition across age groups: using the permuted datasets described above, we extracted the average difference score for the two language conditions (English minus non-English) for each participant, computing an overall average

496 for each random permutation of the data. Then, within each permutation, we  
497 made pairwise comparisons of the average difference scores across participant  
498 age groups. This process yielded a distribution of random permutation-based  
499 difference scores that we could then compare to the difference score in the  
500 actual data. Details are given in Appendix B.

501 These analyses revealed that, while 3- and 4-year olds showed similarly-  
502 sized effects of language condition, 5-year-olds had a significantly smaller  
503 effect of language condition, compared to both younger age groups. The  
504 difference in the language condition effect between 5-year-olds and 3-year-  
505 olds was greater than would be expected by chance (99.52% of the randomly  
506 permuted data sets;  $p < .01$ ). Similarly, the difference in the language con-  
507 dition effect between 5-year-olds and 4-year-olds was greater than would be  
508 expected by chance (99.96% of the data sets;  $p < .001$ ). See Figure B.1 for  
509 each difference score distribution.

510 When does spontaneous turn prediction emerge developmentally? To  
511 test whether the youngest age group (3-year-olds) already exceeded chance  
512 in their anticipatory gaze switches, we compared children's real gaze rates  
513 to the random baseline in the English condition with two-tailed  $t$ -tests.  
514 We used the English condition because we are most interested in finding

515 out when children begin to make spontaneous turn predictions for natural  
516 speech. We found that three-year-olds made anticipatory gaze switches signif-  
517 icantly above chance, when all transitions were considered ( $t(22.824)=-4.147$ ,  
518  $p<.001$ ) as well as for question transitions alone ( $t(21.677)=-5.268$ ,  $p<.001$ ).

519 *2.3. Discussion*

520 Children and adults spontaneously tracked the turn structure of the con-  
521 versations, making anticipatory gaze switches at an above-chance rate across  
522 all ages and conditions. Children's anticipatory gaze rates were affected by  
523 language condition, transition type, age, and gap duration (Table 2), none of  
524 which could be explained by a baseline of random gaze switching (Appendix  
525 A; Figure A.1a). These data show a number of important features that bear  
526 on our questions of interest.

527 First, both adults' and children's anticipations were strongly affected by  
528 transition type. Both groups made more anticipatory switches after hearing  
529 questions, compared to non-questions. Even in the English videos, when  
530 participants had full access to linguistic cues, their rates of anticipation were  
531 relatively low (even comparable to the non-English videos) unless the turn  
532 was a question. Prior work using online, metalinguistic tasks has shown  
533 that participants can use linguistic cues to accurately predict upcoming turn

534 ends (Torreira et al., 2015; Magyari and De Ruiter, 2012; De Ruiter et al.,  
535 2006). The current results add a new dimension to our understanding of  
536 how listeners make predictions about turn ends: both children and adults  
537 spontaneously monitor the linguistic structure of unfolding turns for cues to  
538 imminent responses.

539 Second, children made more anticipatory switches overall in English videos,  
540 compared to non-English videos. This effect suggests that lexical access is  
541 important for children's ability to anticipate upcoming turn structure, con-  
542 sistent with prior work on turn-end prediction in adults (De Ruiter et al.,  
543 2006; Magyari and De Ruiter, 2012) and children (Keitel et al., 2013).

544 Third, we saw that older children made anticipatory switches more re-  
545 liably than younger children, but only in the non-English videos. In the  
546 English videos, children anticipated well at all ages, especially after hear-  
547 ing questions. This interaction between age and language condition suggests  
548 that the 5-year-olds were able to leverage anticipatory cues in the non-English  
549 videos in a way that 3- and 4-year-olds could not, possibly by shifting more  
550 attention to the non-native prosodic or non-verbal cues. Prior work on chil-  
551 dren's turn-structure anticipation has proposed that children's turn-end pre-  
552 dictions rely primarily on lexicosyntactic structure (and not, e.g., prosody)

553 as they get older (Keitel et al., 2013). The current results suggest more  
554 flexibility in children's predictions; when they do not have access to lexical  
555 information, older children and adults are likely to find alternative cues to  
556 turn taking behavior.

557 In Experiment 2, we follow up on these findings, improving on two as-  
558 pects of the design: first, our language manipulation in this first experiment  
559 was too coarse to provide data regarding specific linguistic channels (e.g.,  
560 prosody vs. lexicosyntax). In Experiment 2, we compared lexicosyntactic  
561 and prosodic cues with phonetically altered speech and used puppets to elim-  
562 inate non-verbal cues to turn taking. Second, we were not able to pinpoint  
563 the emergence of anticipatory switching because the youngest age group in  
564 our sample was already able to make anticipatory switches at above chance  
565 rates. In Experiment 2, we explore a wider developmental range.

566 **3. Experiment 2**

567 Experiment 2 used native-language stimuli, controlled for lexical and  
568 prosodic information, eliminating non-verbal cues, and tested children from a  
569 wider age range. To tease apart the role of lexical and prosodic information,  
570 we phonetically manipulated the speech signal for pitch, syllable duration,

571 and lexical access. By testing 1- to 6-year-olds we hoped to find the devel-  
572 opmental onset of turn-predictive gaze. We also hoped to measure changes  
573 in the relative roles of prosody and lexicosyntax across development.

574 Non-verbal cues in Experiment 1 (e.g., gaze and gesture) could have  
575 helped participants make predictions about upcoming turn structure (Rossano  
576 et al., 2009; Stivers and Rossano, 2010). Since our focus was on linguistic  
577 cues, we eliminated all gaze and gestural signals in Experiment 2 by replacing  
578 the videos of human actors with videos of puppets. Puppets are less real-  
579 istic and expressive than human actors, but they create a natural context  
580 for having somewhat motionless talkers in the videos (thereby allowing us  
581 to eliminate gestural and gaze cues). Additionally, the prosody-controlled  
582 condition included small but global changes to syllable duration that would  
583 have required complex video manipulation or precise re-enactment with hu-  
584 man talkers, neither of which was feasible. For these reasons, we decided to  
585 substitute puppet videos for human videos in the final stimuli.

586 As in the first experiment, we recorded participants' eye movements as  
587 they watched six short videos of dyadic conversation, and then analyzed  
588 their anticipatory glances from the current speaker to the upcoming speaker  
589 at points of turn transition.

590    *3.1. Methods*

591    *3.1.1. Participants*

592    We recruited 27 undergraduate adults and 129 children between ages  
593    1;0–6;11 to participate in our experiment. We recruited our child partici-  
594    pants from the Children’s Discovery Museum of San Jose, California<sup>7</sup>, tar-  
595    geting approximately 20 children for each of the six one-year age groups  
596    (range: 20–23). All participants were native English speakers, though some  
597    parents (N=27) reported that their child heard a second (and sometimes  
598    third) language at home. None of the adult participants reported fluency in  
599    a second language.

600    *3.1.2. Materials*

601    We created 18 short videos of improvised, child-friendly conversation (Fig-  
602    ure 5). To eliminate non-verbal cues to turn transition and to control the  
603    types of linguistic information available in the stimuli we first audio-recorded  
604    improvised conversations, then phonetically manipulated those recordings to  
605    limit the availability of prosodic and lexical information, and finally recorded

---

<sup>7</sup>We ran Experiment 2 at a local children’s museum because it gave us access to children with a more diverse range of ages.

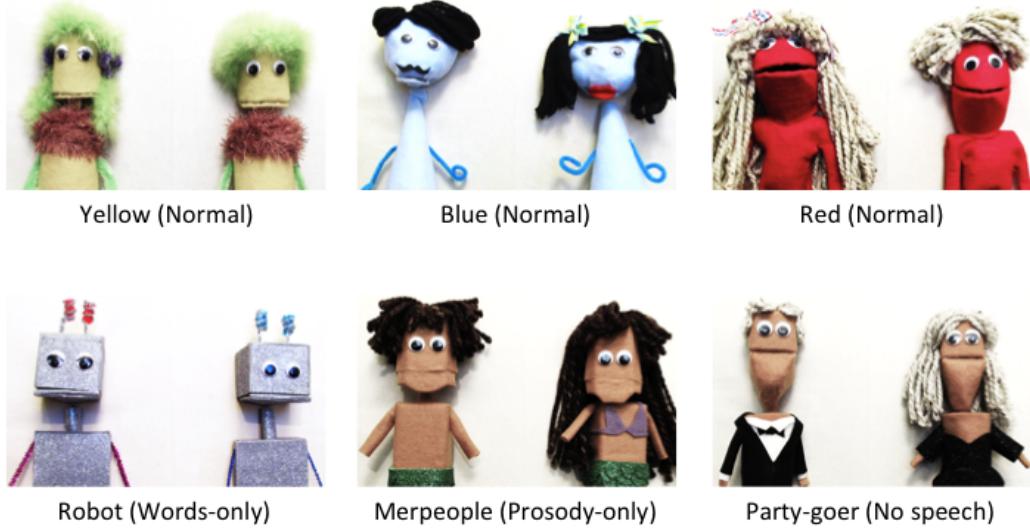


Figure 5: The six puppet pairs (and associated audio conditions). Each pair was linked to three distinct conversations from the same condition across the three experiment versions.

606 video to accompany the manipulated audio, featuring puppets as talkers.  
 607 *Audio recordings.* The recording session was set up in the same way as  
 608 the first experiment, but with a shorter warm up period (5–10 minutes) and  
 609 a pre-determined topic for the child-friendly improvisation ('riding bikes',  
 610 'pets', 'breakfast', 'birthday cake', 'rainy days', or 'the library'). All of the  
 611 talkers were native English speakers, and were recorded in male-female pairs.  
 612 As before, we asked talkers to speak "as if they were on a children's television  
 613 show" and to ask at least a few questions during the improvisation. We cut

614 each audio recording down to the (approximate) 20-second interval with the  
615 most turn activity. The 20-second clips were then phonetically manipulated  
616 and used in the final video stimuli.

617 *Audio Manipulation.* We created four versions of each audio clip: *nor-*  
618 *mal*, *words only*, *prosody only*, and *no speech*. That is, one version with a full  
619 linguistic signal (*normal*), and three with incomplete linguistic information  
620 (hereafter “partial cue” conditions). The *normal* clips were the unmanipu-  
621 lated, original audio clips.

622 The *words only* clips were manipulated to have robot-like speech: we  
623 flattened the intonation contours to each talker’s average pitch ( $F_0$ ) and  
624 we reset the duration of every nucleus and coda to each talker’s average  
625 nucleus and coda duration.<sup>8</sup> We made duration and pitch manipulations  
626 using PSOLA resynthesis in Praat (Boersma and Weenink, 2012). Thus,  
627 the *words only* versions of the audio clips had no pitch or durational cues  
628 to upcoming turn boundaries, but did have intact lexicosyntactic cues (and  
629 residual phonetic correlates of prosody, e.g., intensity).

630 We created the *prosody only* clips by low-pass filtering the original record-

---

<sup>8</sup>We excluded hyper-lengthened words like [wau!] ‘woooow!’. These were rare in the clips.

631 ing at 500 Hz with a 50 Hz Hanning window (following de Ruiter et al., 2006).

632 This manipulation creates a “muffled speech” effect because low-pass filter-

633 ing removes most of the phonetic information used to distinguish between

634 phonemes. The *prosody only* versions of the audio clips lacked lexical infor-

635 mation, but retained their intonational and rhythmic cues to upcoming turn

636 boundaries.

637 The *no speech* condition served as a non-linguistic baseline. For this

638 condition, we replaced the original clip with multi-talker babble: we overlaid

639 multiple child-oriented conversations (not including the original one), and

640 then cropped the result to the duration of the original video. Thus, the

641 *no speech* audio clips lacked any linguistic information to upcoming turn

642 boundaries—the only cue to turn taking was the opening and closing of the

643 puppets’ mouths.

644 Finally, because low-pass filtering removes significant acoustic energy, the

645 *prosody only* clips were much quieter than the other three conditions. Our

646 last step was to downscale the intensity of the audio tracks in the three other

647 conditions to match the volume of the *prosody only* clips. We referred to the

648 conditions as “normal”, “robot”, “mermaid”, and “birthday party” speech

649 when interacting with participants.

650        *Video recordings.* We created puppet video recordings to match the ma-  
651        nipulated 20-second audio clips. The puppets were minimally expressive;  
652        so that the puppeteer could only control the opening and closing of their  
653        mouths; the puppets' heads, eyes, arms, and bodies stayed still. Puppets  
654        were positioned looking forward to eliminate shared gaze as a cue to turn  
655        structure (Thorgrímsson et al., 2015). We took care to match the puppets'  
656        mouth movements to the syllable onsets as closely as possible, specifically  
657        avoiding any mouth movement before the onset of a turn. We then added  
658        the manipulated audio clips to the puppet video recordings by hand.

659        We used three pairs of puppets used for the *normal* condition—‘red’,  
660        ‘blue’ and ‘yellow’—and one pair of puppets for each partial cue condition:  
661        “robots”, “merpeople”, and “party-goers” (Figure 8). We randomly assigned  
662        half of the conversation topics (‘birthday cake’, ‘pets’, and ‘breakfast’) to the  
663        *normal* condition, and half to the partial cue conditions (‘riding bikes’, ‘rainy  
664        days’, and ‘the library’). We then created three versions of the experiment,  
665        so that each of the six puppet pairs was associated with three different con-  
666        versation topics across the different versions of the experiment (18 videos  
667        in total). We ensured that the position of the talkers (left and right) was  
668        counterbalanced in each version by flipping the video and audio channels as

669 needed.

670 The duration of turn transitions and the number of speaker changes  
671 across videos was variable because the conversations were recorded semi-  
672 spontaneously. We measured turn transitions from the audio signal of the  
673 *normal*, *words only*, and *prosody only* conditions. There was no audio from  
674 the original conversation in the *no speech* condition videos, so we measured  
675 turn transitions from puppets' mouth movements in the video signal, using  
676 ELAN video annotation software (Wittenburg et al., 2006).

677 There were 85 turn transitions for analysis after excluding transitions  
678 longer than 550 msec and shorter than 90 msec. The remaining turn transi-  
679 tions had more questions than non-questions (N=47 and N=38, respectively),  
680 with transitions distributed somewhat evenly across conditions (keeping in  
681 mind that there were three *normal* videos and only one partial cue video for  
682 each experiment version): *normal* (N=36), *words only* (N=13), *prosody only*  
683 (N=17), and *no speech* (N=19). Inter-turn gaps for questions (mean=366,  
684 median=438, stdev=138 msec) were longer than those for non-questions  
685 (mean=305, median=325, stdev=94 msec) on average, but gap duration  
686 was overall comparable across conditions: *normal* (mean=334, median=321,  
687 stdev=130 msec), *words only* (mean=347, median=369, stdev= 115 msec),

688 *prosody only* (mean=365, median=369, stdev=104 msec), and *no words*  
689 (mean=319, median=329, stdev=136 msec). The longer gaps for question  
690 transitions could give them an advantage because our anticipatory measure  
691 includes shifts initiated during the gap between turns (Figure 2).

### 692 3.2. Procedure

693 We used the same experimental apparatus and procedure as in the first  
694 experiment. Each participant watched six puppet videos in random order,  
695 with five 15–30 second filler videos placed in-between (e.g., running puppies,  
696 moving balls, flying bugs). Three of the puppet videos had *normal* audio  
697 while the other three had *words only*, *prosody only*, and *no speech* audio.  
698 This experiment required no special instructions so, as before, the experi-  
699 menter immediately began each session with calibration and then stimulus  
700 presentation. The entire experiment took less than five minutes.

#### 701 3.2.1. Data preparation and coding

702 We coded each turn transition for its linguistic condition (*normal*, *words*  
703 *only*, *prosody only*, and *no speech*) and transition type (question/non-question)<sup>9</sup>,

---

<sup>9</sup>We coded *wh*-questions as “non-questions” for the *prosody only* videos. Polar questions often have a final rising intonational contour, but *wh*-questions often do not (Hedberg et al.,

Age group	Speaker	Addressee	Other onscreen	Offscreen
1	0.44	0.14	0.23	0.19
2	0.50	0.13	0.24	0.14
3	0.47	0.12	0.25	0.16
4	0.48	0.11	0.29	0.12
5	0.54	0.11	0.20	0.14
6	0.60	0.12	0.18	0.10
Adult	0.69	0.12	0.09	0.10

Table 3: Average proportion of gaze to the current speaker and addressee during periods of talk across ages.

Condition	Speaker	Addressee	Other onscreen	Offscreen
Normal	0.58	0.12	0.17	0.13
Words only	0.54	0.11	0.24	0.10
Prosody only	0.48	0.12	0.26	0.15
No speech	0.44	0.13	0.26	0.18

Table 4: Average proportion of gaze to the current speaker and addressee during periods of talk across conditions.

704 and identified anticipatory gaze switches to the upcoming speaker using the  
 705 methods from Experiment 1.

706 *3.3. Results*

707 Participants' pattern of gaze indicated that they performed basic turn  
 708 tracking across all ages and in all conditions. Participants looked at the  
 709 screen most of the time during video playback (82% and 86% average for

---

2010).

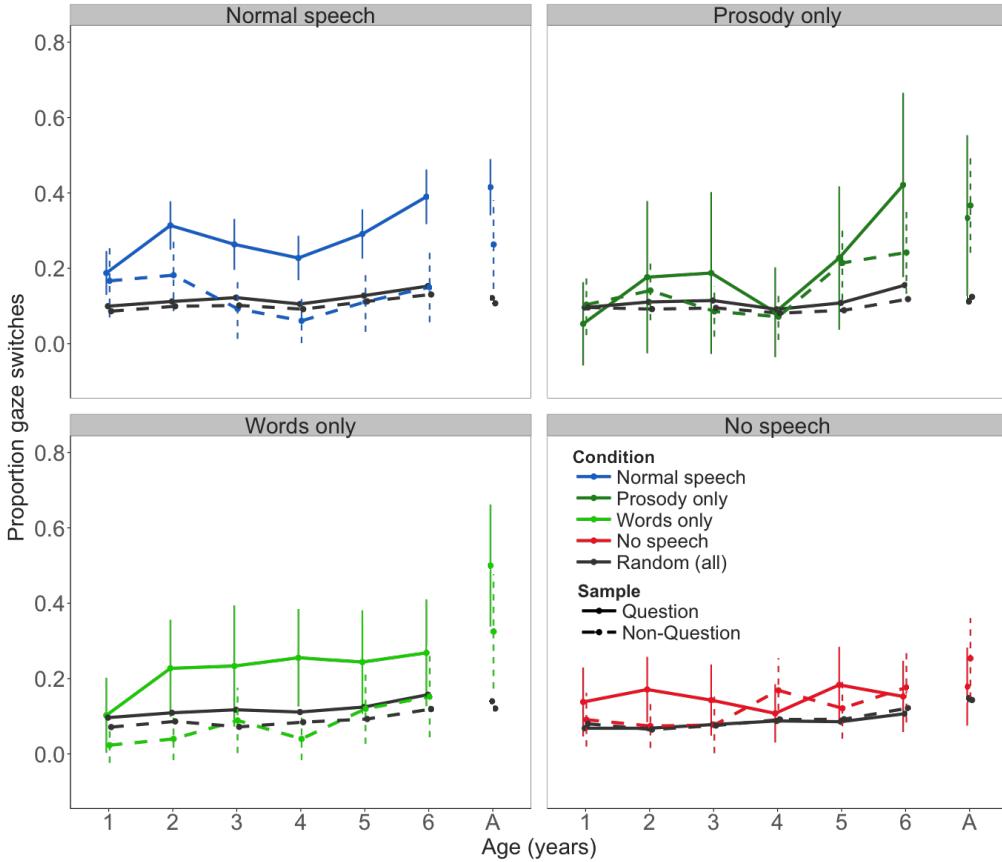


Figure 6: Anticipatory gaze rates across language condition and transition type for the real data (blue, dark green, light green, and red) and the randomly permuted baselines (gray). Vertical bars represent 95% confidence intervals.

<sup>710</sup> children and adults, respectively), primarily looking at the person who was  
<sup>711</sup> currently speaking (Table 2). They tracked the current speaker in every

712 condition—even one-year-olds looked more at the current speaker than at  
713 anything else in the three partial cue conditions (40% for *words only*, 43%  
714 for *prosody only*, and 39% for *no speech*). There was a steady overall increase  
715 in looks to the current speaker with age and added linguistic information  
716 (Tables 3 and 4). Looks to the addressee also decreased with age, but the  
717 change was minimal. Figure 6 shows participants' anticipatory gaze rates  
718 across age, the four language conditions, and transition type.

719 *3.3.1. Statistical models*

720 We identified anticipatory gaze switches for all 85 usable turn transi-  
721 tions, and analyzed them for effects of language condition, transition type,  
722 and age with two mixed-effects logistic regressions. We again built separate  
723 models for children and adults because effects of age were only pertinent to  
724 the children's data. The child model included condition (*normal/prosody*  
725 *only/words only/no speech*; with *no speech* as the reference level), transition  
726 type (question vs. non-question), age (1, 2, 3, 4, 5, 6; numeric), and duration  
727 of the inter-turn gap (in seconds) as predictors, with full interactions between  
728 language condition, transition type, and age. We again included the dura-  
729 tion of the inter-turn gap as a control predictor and added random effects of  
730 item (turn transition) and participant, with random slopes of transition type

731 for participants. The adult model included condition, transition type, their  
732 interactions, and duration as a control predictor, with participant and item  
733 included as random effects and random slopes of condition and transition  
734 type.

735 Children's anticipatory gaze switches showed an effect of gap duration  
736 ( $\beta=3.95$ ,  $SE=0.617$ ,  $z=6.401$ ,  $p<.001$ ), a two-way interaction of age and  
737 language condition (for *prosody only* speech compared to the *no speech* refer-  
738 ence level;  $\beta=0.393$ ,  $SE=0.189$ ,  $z=2.08$ ,  $p<.05$ ), and a three-way interaction  
739 of age, transition type, and language condition (for *normal* speech compared  
740 to the *no speech* reference level;  $\beta=-0.38$ ,  $SE=0.17$ ,  $z=-2.229$ ,  $p<.05$ ). There  
741 were no significant effects of age or transition type alone (Table 3.3.1), with  
742 only a marginal effect of language condition (for *prosody only* compared to  
743 the *no speech* reference level;  $\beta=-1.607$ ,  $SE=0.867$ ,  $z=-1.85$ ,  $p=.064$ )

744 Adults' anticipatory gaze switches showed effects of gap duration ( $\beta=4.7$ ,  
745  $SE=1.18$ ,  $z=3.978$ ,  $p<.001$ ) and language condition (for *normal* speech  $\beta=1.203$ ,  
746  $SE=0.536$ ,  $z=2.245$ ,  $p<.05$  and *words only* speech  $\beta=1.561$ ,  $SE=0.709$ ,  $z=2.203$ ,  
747  $p<.05$  compared to the *no speech* reference level). There were no effects of  
748 transition type ( $\beta=0.437$ ,  $SE=0.585$ ,  $z=0.747$ ,  $p=.45$ ).

749 *3.3.2. Random baseline comparison*

750 Using the same technique described in Experiment 1 (Section 2.2.2), we  
751 created and modeled random permutations of participants' anticipatory gaze.  
752 These analyses revealed that none of the significant predictors from models  
753 of the original, turn-related data could be explained by random looking.  
754 In the children's data, the original model's *z*-values for language condition  
755 (*prosody only*), gap duration, the two-way interaction of age and language  
756 condition (*prosody only*) and the three-way interaction of age, transition type,  
757 and language condition (*normal speech*) were all greater than 95% of the  
758 randomly permuted *z*-values (96.5%, 100%, 95.6%, and 95.5%, respectively,  
759 all  $p < .05$ ). Similarly, the adults' data showed significant differentiation from  
760 the randomly permuted data for all originally significant predictors: gap  
761 duration and language condition for *normal speech* and *words only* speech  
762 (greater than 100%, 97.8%, and 99.1% of random *z*-values, respectively, all  
763  $p < .05$ ). See Section Appendix A for more information on each predictor's  
764 random permutation distribution.

<i>Children</i>	Estimate	Std. Error	<i>z</i> value	Pr(>  <i>z</i>  )
(Intercept)	-3.49403	0.48454	-7.211	5.55e-13 ***
Age	0.02436	0.10249	0.238	0.8121
Type= <i>non-Question</i>	-0.88900	0.61192	-1.453	0.1463
GapDuration	3.94743	0.61668	6.401	1.54e-10 ***
Age*Type= <i>non-Question</i>	0.15359	0.13996	1.097	0.2725
Condition= <i>normal</i>	0.37337	0.43421	0.860	0.3899
Age*Condition= <i>normal</i>	0.12950	0.10217	1.267	0.2050
Condition= <i>normal</i> *	0.91074	0.72581	1.255	0.2096
Type= <i>non-Question</i>				
Age*Condition= <i>normal</i> *	-0.37965	0.17031	-2.229	0.0258 *
Type= <i>non-Question</i>				
Condition= <i>prosody</i>	-1.60734	0.86680	-1.854	0.0637 .
Age*Condition= <i>prosody</i>	0.39271	0.18905	2.077	0.0378 *
Condition= <i>prosody</i> *	1.68552	1.05414	1.599	0.1098
Type= <i>non-Question</i>				
Age*Condition= <i>prosody</i> *	-0.32360	0.23229	-1.393	0.1636
Type= <i>non-Question</i>				
Condition= <i>words</i>	-0.26996	0.59313	-0.455	0.6490
Age*Condition= <i>words</i>	0.14044	0.13565	1.035	0.3005
Condition= <i>words</i> *	-1.03066	1.01610	-1.014	0.3104
Type= <i>non-Question</i>				
Age*Condition= <i>words</i> *	0.08829	0.22387	0.394	0.6933
Type= <i>non-Question</i>				
<i>Adults</i>	Estimate	Std. Error	<i>z</i> value	Pr(>  <i>z</i>  )
(Intercept)	-3.3811	0.6884	-4.912	9.03e-07 ***
Type= <i>non-Question</i>	0.4375	0.5854	0.747	0.4549
GapDuration	4.6961	1.1804	3.978	6.94e-05 ***
Condition= <i>normal</i>	1.2033	0.5359	2.245	0.0247 *
Condition= <i>normal</i> *	-0.9627	0.7358	-1.308	0.1907
Type= <i>non-Question</i>				
Condition= <i>prosody</i>	0.2407	0.8011	0.301	0.7638
Condition= <i>prosody</i> *	0.5525	0.9374	0.589	0.5556
Type= <i>non-Question</i>				
Condition= <i>words</i>	1.5613	0.7087	2.203	0.0276 *
Condition= <i>words</i> *	-1.1557	0.8854	-1.305	0.1918
Type= <i>non-Question</i>				

Table 5: Model output for children and adults' anticipatory gaze switches in Experiment 2.

766 3.3.3. Developmental effects

767 Our main goal in extending the age range to 1- and 2-year-olds in Ex-  
768 periment 2 was to find the age of emergence for spontaneous predictions  
769 about upcoming turn structure. As in Experiment 1, we used two-tailed  
770 *t*-tests to compare children's real gaze rates to the random baseline rates  
771 in the *normal* speech condition, in which the speech stimulus is most like  
772 what children hear every day. We tested real gaze rates against baseline for  
773 three age groups: ages one, two, and three. Two- and three-year-old children  
774 made anticipatory gaze switches significantly above chance both when all  
775 transitions were considered (2-year-olds:  $t(26.193)=-4.137$ ,  $p<.001$ ; 3-year-  
776 olds:  $t(22.757)=-2.662$ ,  $p<.05$ ) and for question transitions alone (2-year-  
777 olds:  $t(25.345)=-4.269$ ,  $p<.001$ ; 3-year-olds:  $t(21.555)=-3.03$ ,  $p<.01$ ). One-  
778 year-olds, however, only made anticipatory gaze shifts marginally above  
779 chance for turn transitions overall and for question turns alone (overall:  
780  $t(24.784)=-2.049$ ,  $p=.051$ ; questions:  $t(25.009)=-2.03$ ,  $p=.053$ ).

781 The regression models for the children's data also revealed two signifi-  
782 cant interactions with age. The first was a significant interaction of age and  
783 language condition (for *prosody only* compared to the *no speech* reference  
784 level), suggesting a different age effect between the two linguistic conditions.

785 As in Experiment 1, we explored each age interaction by extracting an av-  
786 erage difference score over participants for the effect of language condition  
787 (*no speech* vs. *prosody only*) within each random permutation of the data,  
788 making pairwise comparisons between the six age groups. These tests re-  
789 vealed that children’s anticipation in the *prosody only* condition significantly  
790 improved at ages five and six (with difference scores greater than 95% of the  
791 random data scores;  $p < .05$ ). See Figure B.2 for these *prosody only* difference  
792 score distributions.

793 The second age-based interaction was a three-way interaction of age, tran-  
794 sition type, and language condition (for *normal* speech compared to the *no*  
795 *speech* baseline). We again created pairwise comparisons of the average dif-  
796 ference scores for the transition type-language condition interaction across  
797 age groups in each random permutation of the data, finding that the effect  
798 of transition type in the *normal* speech condition became larger with age,  
799 with significant improvements by age 4 over ages 1 and 2 (99.9% and 98.86%,  
800 respectively), by age 5 over age 4 (97.54%), and by age 6 over ages 1, 2, and 5  
801 (99.5%, 97.36%, and 95.04%), all significantly different from chance ( $p < .05$ ).  
802 See Figure B.3 for these *normal* speech difference score distributions.

803    3.4. Discussion

804    The core aims of Experiment 2 were to gain better traction on the individual roles of prosody and lexicosyntax in children’s turn predictions, and to  
805    find the age of emergence for spontaneous turn anticipation. Many of our results replicate the findings from Experiment 1: participants often made more  
806    anticipatory switches when they had access to lexical information and, when  
807    they did, tended to make more anticipatory switches for questions compared  
808    to non-questions.

811    As in Experiment 1, children and adults spontaneously tracked the turn  
812    structure of the conversations, making anticipatory gaze switches at above-chance rates across all ages when listening to natural speech. They also made  
813    far more anticipations for questions than for non-question turns—at least for  
814    children two years old and older. But these effects were different for the two  
815    conditions with partial linguistic information: *prosody only* and *words only*.  
817    In the *prosody only* condition, performance was low for younger children and  
818    increased significantly with age (especially for questions). In the *words only*  
819    condition, children age two and older showed robust anticipatory switching  
820    for questions (much like in *normal* speech), but never rose above chance for  
821    non-question turns. These findings do not therefore support an early role

822 for prosody in children's spontaneous turn structure predictions. There is  
823 also no evidence that lexical information is sufficient on its own to support  
824 children's anticipatory switching.

825 **4. General Discussion**

826 Children begin to develop conversational turn-taking skills long before  
827 their first words emerge (Bateson, 1975; Hilbrink et al., 2015; Jaffe et al.,  
828 2001; Snow, 1977). As they acquire language, they also acquire the infor-  
829 mation needed to make accurate predictions about upcoming turn structure.  
830 Until recently, we have had very little data on how children weave language  
831 into their already-existing turn-taking behaviors. In two experiments inves-  
832 tigating children's anticipatory gaze to upcoming speakers, we found evi-  
833 dence that turn prediction develops early in childhood and that spontaneous  
834 predictions are primarily driven by participants' expectation of an imme-  
835 diate response in the next turn (e.g., after questions). In making predic-  
836 tions about upcoming turn structure, children used a combination of lexical  
837 and prosodic cues; neither lexical nor prosodic cues alone were sufficient to  
838 support increased anticipatory gaze. We also found no early advantage for  
839 prosody over lexicosyntax, and instead found that children were unable to

840 make above-chance anticipatory gazes in the *prosody only* condition until age  
841 five. We discuss these findings with respect to the role of linguistic cues in  
842 predictions about upcoming turn structure, the importance of questions in  
843 spontaneous predictions about conversation, and children's developing com-  
844 petence as conversationalists.

845 *4.1. Predicting turn structure with linguistic cues*

846 Prior work with adults has found a consistent role for lexicosyntax in  
847 predicting upcoming turn structure (De Ruiter et al., 2006; Magyari and  
848 De Ruiter, 2012), whereas the role of prosody still under debate (Duncan,  
849 1972; Ford and Thompson, 1996; Torreira et al., 2015). Knowing that chil-  
850 dren comprehend more about prosody than lexicosyntax early on (Section  
851 1; also see Speer and Ito, 2009 for a review), we thought it possible that  
852 young children would instead show an advantage for prosody in their predic-  
853 tions about turn structure in conversation. Our results suggest that, on the  
854 contrary, exclusively presenting prosodic information to children limits their  
855 spontaneous predictions about upcoming turn structure until age five.

856 Perhaps surprisingly, we also found no evidence that lexical information  
857 alone is equivalent to the full linguistic signal in driving children's predic-  
858 tions, as has been shown previously for adults (Magyari and De Ruiter, 2012;

859 De Ruiter et al., 2006) and as is replicated with adult participants in the cur-  
860 rent study. That said, our findings point more toward early lexical effects  
861 in children’s turn anticipations than prosodic ones: children’s performance  
862 in both experiments was consistently better when they had access to lexical  
863 information, especially after question turns. And although the *words only*  
864 condition in Experiment 2 was not significantly different from the baseline *no*  
865 *speech* condition, children’s anticipations trended toward above-chance rates.

866 Above all, children and adults anticipated best with they had access to  
867 the full linguistic signal. There may be something informative about com-  
868 bined prosodic and lexical cues to questionhood that helps to boost children’s  
869 anticipations before they can use these cues separately. Even in adults, Tor-  
870 reira and colleagues (2015) showed that the trade-off in informativity between  
871 lexical and prosodic cues is more subtle in semi-natural speech. The present  
872 findings are the first to show evidence of a similar effect developmentally.

873 *4.2. The question effect*

874 In both experiments, anticipatory looking was primarily driven by ques-  
875 tion transitions, a pattern that has not been previously reported in other an-  
876 ticipatory gaze studies, on children or adults (Keitel et al., 2013; Hirvenkari,  
877 2013; Tice and Henetz, 2011). Questions make an upcoming speaker switch

878 immediately relevant, helping the listener to predict with high certainty what  
879 will happen next (i.e., an answer from the addressee), and are often easily  
880 identifiable by overt prosodic and lexicosyntactic cues.

881 Compared to prosodic cues (e.g., final rising intonation), lexicosyntactic  
882 cues to questionhood (e.g., *wh*-words, *do*-insertion, and subject-auxiliary in-  
883 version) are categorical, and early-occurring in the utterance. Children may  
884 have therefore had an easier time picking out and interpreting lexical cues  
885 to questionhood. Developmentally, the question effect showed its first signif-  
886 icant gains between ages three and four in the *normal* speech condition of  
887 Experiment 2 (Figure B.3), by which time children frequently hear and use  
888 a variety of polar and *wh*-questions (Clark, 2009). Furthermore, while lexi-  
889 cosyntactic question cues were available on every instance of *wh*- and *yes/no*  
890 questions in our stimuli, prosodic question cues were only salient on *yes/no*  
891 questions. Finally, the mapping of prosodic contour to speech act (e.g., high  
892 final rises for polar questions) is far from one-to-one, leaving substantial room  
893 for uncertainty in prosodic contour interpretation in general.

894 Prior work on children's acquisition of questions indicates that they may  
895 already have some knowledge of question-answer sequences by the time they  
896 begin to speak: questions make up approximately one third of the utter-

897 ances children hear, before and after the onset of speech, and even into  
898 their preschool years, though the type and complexity of questions changes  
899 throughout development (Casillas et al., In press; Fitneva, 2012; Henning  
900 et al., 2005; Shatz, 1979).<sup>10</sup> For the first few years, many of the questions  
901 directed to children are “test” questions—questions that the caregiver al-  
902 ready has the answer to (e.g., “What does a cat say?”), but this changes as  
903 children get older. Questions help caregivers to get their young children’s  
904 attention and to ensure that information is in common ground, even if the  
905 responses are non-verbal or infelicitous (Bruner, 1985; Fitneva, 2012; Snow,  
906 1977). So, in addition to having a special interactive status (for adults and  
907 children alike), questions are a core characteristic of many caregiver-child  
908 interactions, motivating a general benefit for questions in turn structure an-  
909 ticipation.

910 Two important questions for future work are then: (1) how does children’s  
911 ability to monitor for questions in conversation relate to their prior experience  
912 with questions? and (2) what is it about questions that makes children and  
913 adults more likely to anticipatorily switch their gaze to addressees? Other

---

<sup>10</sup>There is substantial variation in question frequency by individual and socioeconomic class (Hart and Risley, 1992; Weisleder, 2012).

914 turn types, such as imperatives, compliments, and complaints make a re-  
915 sponse from the addressee highly likely in the next turn (Schegloff, 2007).  
916 Rhetorical and tag questions, on the other hand, take a similar form to pro-  
917 totypical polar questions, but often do not require an answer. So, though it  
918 is clear that adults and children anticipated responses more often for ques-  
919 tions than non-questions, we do not yet know whether their predictive action  
920 is limited to turns formatted as questions or is generally applicable to turn  
921 structures that project an immediate response from the addressee.

922 More broadly, our results suggest that participants' spontaneous predic-  
923 tions, at least while viewing third-party conversation, are driven by what lies  
924 *beyond* the end of the current turn—not just by the upcoming end of the  
925 turn itself, as has been focused on in prior work (Torreira et al., 2015; Keitel  
926 et al., 2013; Magyari and De Ruiter, 2012; De Ruiter et al., 2006). In future  
927 work, it will be crucial to measure prediction from a first-person perspective  
928 to resolve this apparent contradiction (see also Holler and Kendrick, 2015).

929 *4.3. Early competence for turn taking?*

930 One of the core aims of our study was to test whether children show an  
931 early competence for turn taking, as is proposed by studies of spontaneous  
932 mother-infant proto-conversation and theories about the mechanisms under-

933 lying human interaction in general (Hilbrink et al., 2015; Levinson, 2006).

934 We found evidence that young children make spontaneous predictions about

935 upcoming turn structure: definitely at age two and marginally at age one.

936 These results contrast with Keitel and colleagues' (2013) finding that chil-

937 dren cannot anticipate upcoming turn structure at above-chance rates until

938 age three. The current study used an appreciably more conservative random

939 baseline than the one used in Keitel and colleagues' study. Therefore, this

940 difference in age of emergence more likely stems from our use of a more en-

941 gaging speech style, stereo speech playback, and more typical turn transition

942 durations.

943 To be clear, young children's "above chance" performance was often still

944 far from adult-like predictive behavior—children at ages one and two were

945 still very close to chance in their anticipations and, even at age six, children

946 were not fully adult-like in their predictions. This may indicate that young

947 children rely more on non-verbal cues in anticipating turn transitions or,

948 alternatively, that adults are better at flexibly adapting to the turn-relevant

949 cues present at any moment.

950 Taken together, our data suggest that turn-taking skills do begin to

951 emerge in infancy, but that children cannot make effective predictions un-

952 til they can pick out question turns. This finding leads us to wonder how  
953 participant role (first- instead of third-person) and cultural differences (e.g.,  
954 high vs. low parent-infant interaction styles) feed into this early predictive  
955 skill. It also bridges prior work showing a predisposition for turn taking in  
956 infancy (e.g., Bateson, 1975; Hilbrink et al., 2015; Jaffe et al., 2001; Snow,  
957 1977) with children's apparently *late* acquisition of adult-like competence for  
958 turn taking in actual conversation (Casillas et al., In press; Garvey, 1984;  
959 Garvey and Berninger, 1981; Ervin-Tripp, 1979).

960 *4.4. Limitations and future work*

961 There are at least two major limitations to our work: speech naturalness  
962 and participant role. Following prior work (De Ruiter et al., 2006; Keitel  
963 et al., 2013), we used phonetically manipulated speech in Experiment 2.  
964 This decision resulted in speech sounds that children don't usually hear in  
965 their natural environment. Many prior studies have used phonetically-altered  
966 speech with infants and young children (cf. Jusczyk, 2000), but almost none  
967 of them have done so in a conversational context. Future work could instead  
968 carefully script or cross-splice sub-parts of turns to control for the presence  
969 of linguistic cues for turn transition (see, e.g., Torreira et al., 2015).

970 The prediction measure used in our studies is based on an observer's view

971 of third-party conversation but, because participants' role in the interaction  
972 could affect their online predictions about turn taking, an better measure  
973 would instead capture first-person behavior. First-person measures of spon-  
974 taneous turn prediction will be key to revealing how participants distribute  
975 their attention over linguistic and non-linguistic cues while taking part in  
976 everyday interaction, the implications of which relate to theories of online  
977 language processing for both language learning and everyday talk.

978 *4.5. Conclusions*

979 Conversation plays a central role in children's language learning. It is  
980 the driving force behind what children say and what they hear. Adults use  
981 linguistic information to accurately predict turn structure in conversation,  
982 which facilitates their online comprehension and allows them to respond rel-  
983 evantly and on time. The present study offers new findings regarding the  
984 role of speech acts and linguistic processing in online turn prediction, and  
985 has given evidence that turn prediction emerges by age two, but is not inte-  
986 grated with linguistic cues until much later. Using language to make predic-  
987 tions about upcoming interactive content takes time and, for both children  
988 and adults, is primarily driven by participants' orientation to what will hap-  
989 pen next—beyond the end of the current turn.

990 **Acknowledgements**

991 We gratefully acknowledge the parents and children at Bing Nursery  
992 School and the Children’s Discovery Museum of San Jose. This work was sup-  
993 ported by an ERC Advanced Grant to Stephen C. Levinson (269484-INTERACT),  
994 an NSF Graduate Research Fellowship and NSF Dissertation Improvement  
995 Grant to MC, and a Merck Foundation fellowship to MCF. Earlier versions  
996 of these data and analyses were presented to conference audiences (Casil-  
997 las and Frank, 2012, 2013). We also thank Tania Henetz, Francisco Tor-  
998 reira, Stephen C. Levinson, Eve V. Clark, and the First Language Acquisi-  
999 tion group at Radboud University for their feedback on earlier versions of  
1000 this work. The analysis code for this project can be found on GitHub at  
1001 [https://github.com/langcog/turn\\_taking/](https://github.com/langcog/turn_taking/).

1002 **References**

1003 Allison, P.D., 2004. Convergence problems in logistic regression, in: Alt-  
1004 man, M., Gill, J., McDonald, M. (Eds.), Numerical Issues in Statistical  
1005 Computing for the Social Scientist. Wiley-Interscience: New York, NY,  
1006 pp. 247–262.

- 1007 Allison, P.D., 2012. Logistic Regression Using SAS: Theory and Application.
- 1008 SAS Institute.
- 1009 Barr, D.J., Levy, R., Scheepers, C., Tily, H.J., 2013. Random effects structure
- 1010 for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory*
- 1011 and Language
- 1012 Bates, D., Maechler, M., Bolker, B., Walker, S., 2014. lme4:
- 1013 Linear mixed-effects models using Eigen and S4. URL:
- 1014 <https://github.com/lme4/lme4>/<http://lme4.r-forge.r-project.org/>.
- 1015 [Computer program] R package version 1.1-7.
- 1016 Bateson, M.C., 1975. Mother-infant exchanges: The epigenesis of conver-
- 1017 sational interaction. *Annals of the New York Academy of Sciences* 263,
- 1018 101–113.
- 1019 Bergelson, E., Swingley, D., 2013. The acquisition of abstract words by young
- 1020 infants. *Cognition* 127, 391–397.
- 1021 Bloom, K., 1988. Quality of adult vocalizations affects the quality of infant
- 1022 vocalizations. *Journal of Child Language* 15, 469–480.

- 1023 Boersma, P., Weenink, D., 2012. Praat: doing phonetics by computer. URL:  
1024 <http://www.praat.org>. [Computer program] Version 5.3.16.
- 1025 Bögels, S., Magyari, L., Levinson, S.C., 2015. Neural signatures of response  
1026 planning occur midway through an incoming question in conversation. Sci-  
1027 entific Reports 5.
- 1028 Bruner, J., 1985. Child's talk: Learning to use language. Child Language  
1029 Teaching and Therapy 1, 111–114.
- 1030 Bruner, J.S., 1975. The ontogenesis of speech acts. Journal of Child Language  
1031 2, 1–19.
- 1032 Carlson, R., Hirschberg, J., Swerts, M., 2005. Cues to upcoming swedish  
1033 prosodic boundaries: Subjective judgment studies and acoustic correlates.  
1034 Speech Communication 46, 326–333.
- 1035 Casillas, M., Bobb, S.C., Clark, E.V., In press. Turn taking, timing, and  
1036 planning in early language acquisition. Journal of Child Language .
- 1037 Casillas, M., Frank, M.C., 2012. Cues to turn boundary prediction in adults  
1038 and preschoolers, in: Proceedings of SemDial, pp. 61–69.
- 1039 Casillas, M., Frank, M.C., 2013. The development of predictive processes

- 1040        in children's discourse understanding, in: Proceedings of the 35th Annual  
1041        Meeting of the Cognitive Science Society, pp. 299–304.
- 1042        Clark, E.V., 2009. First language acquisition. Cambridge University Press.
- 1043        De Ruiter, J.P., Mitterer, H., Enfield, N.J., 2006. Projecting the end of  
1044        a speaker's turn: A cognitive cornerstone of conversation. Language 82,  
1045        515–535.
- 1046        De Vos, C., Torreira, F., Levinson, S.C., 2015. Turn-timing in signed con-  
1047        versations: coordinating stroke-to-stroke turn boundaries. Frontiers in  
1048        Psychology 6.
- 1049        Dingemanse, M., Torreira, F., Enfield, N., 2013. Is “Huh?” a universal word?  
1050        Conversational infrastructure and the convergent evolution of linguistic  
1051        items. PloS one 8, e78273.
- 1052        Duncan, S., 1972. Some signals and rules for taking speaking turns in con-  
1053        versations. Journal of Personality and Social Psychology 23, 283.
- 1054        Ervin-Tripp, S., 1979. Children's verbal turn-taking, in: Ochs, E., Schieffelin,  
1055        B.B. (Eds.), Developmental Pragmatics. Academic Press, New York, pp.  
1056        391–414.

- 1057 Fitneva, S., 2012. Beyond answers: questions and children's learning, in:
- 1058 De Ruiter, J.P. (Ed.), Questions: Formal, Functional, and Interactional
- 1059 Perspectives. Cambridge University Press, Cambridge, UK, pp. 165–178.
- 1060 Ford, C.E., Thompson, S.A., 1996. Interactional units in conversation: Syn-
- 1061 tactic, intonational, and pragmatic resources for the management of turns.
- 1062 Studies in Interactional Sociolinguistics 13, 134–184.
- 1063 Garvey, C., 1984. Children's Talk. volume 21. Harvard University Press.
- 1064 Garvey, C., Berninger, G., 1981. Timing and turn taking in children's con-
- 1065 versations 1. Discourse Processes 4, 27–57.
- 1066 Gísladóttir, R., Chwilla, D., Levinson, S.C., 2015. Conversation electrified:
- 1067 ERP correlates of speech act recognition in underspecified utterances. PloS
- 1068 one 10, e0120068.
- 1069 Griffin, Z.M., Bock, K., 2000. What the eyes say about speaking. Psycho-
- 1070 logical science 11, 274–279.
- 1071 Hart, B., Risley, T.R., 1992. American parenting of language-learning chil-
- 1072 dren: Persisting differences in family-child interactions observed in natural
- 1073 home environments. Developmental Psychology 28, 1096.

- 1074 Hedberg, N., Sosa, J.M., Görgülü, E., Mameni, M., 2010. The prosody and  
1075 meaning of Wh-questions in American English, in: Speech Prosody 2010,  
1076 pp. 100045:1–4.
- 1077 Henning, A., Striano, T., Lieven, E.V., 2005. Maternal speech to infants at  
1078 1 and 3 months of age. *Infant Behavior and Development* 28, 519–536.
- 1079 Hilbrink, E., Gattis, M., Levinson, S.C., 2015. Early developmental changes  
1080 in the timing of turn-taking: A longitudinal study of mother-infant inter-  
1081 action. *Frontiers in Psychology* 6.
- 1082 Hirvenkari, L., Ruusuvuori, J., Saarinen, V.M., Kivioja, M., Peräkylä, A.,  
1083 Hari, R., 2013. Influence of turn-taking in a two-person conversation on  
1084 the gaze of a viewer. *PloS one* 8, e71569.
- 1085 Holler, J., Kendrick, K.H., 2015. Unaddressed participants' gaze in multi-  
1086 person interaction. *Frontiers in Psychology* 6.
- 1087 Jaffe, J., Beebe, B., Feldstein, S., Crown, C.L., Jasnow, M.D., Rochat, P.,  
1088 Stern, D.N., 2001. Rhythms of dialogue in infancy: Coordinated timing in  
1089 development. *Monographs of the Society for Research in Child Develop-  
1090 ment*. JSTOR.

- 1091 Johnson, E.K., Jusczyk, P.W., 2001. Word segmentation by 8-month-olds:  
1092 When speech cues count more than statistics. *Journal of Memory and*  
1093 *Language* 44, 548–567.
- 1094 Jusczyk, P.W., 2000. *The Discovery of Spoken Language*. MIT press.
- 1095 Jusczyk, P.W., Hohne, E., Mandel, D., Strange, W., 1995. Picking up reg-  
1096 ularities in the sound structure of the native language. *Speech perception*  
1097 and linguistic experience: Theoretical and methodological issues in cross-  
1098 language speech research , 91–119.
- 1099 Kamide, Y., Altmann, G., Haywood, S.L., 2003. The time-course of predic-  
1100 tion in incremental sentence processing: Evidence from anticipatory eye  
1101 movements. *Journal of Memory and Language* 49, 133–156.
- 1102 Keitel, A., Daum, M.M., 2015. The use of intonation for turn anticipation  
1103 in observed conversations without visual signals as source of information.  
1104 *Frontiers in Psychology* 6.
- 1105 Keitel, A., Prinz, W., Friederici, A.D., Hofsten, C.v., Daum, M.M., 2013.  
1106 Perception of conversations: The importance of semantics and intonation  
1107 in childrens development. *Journal of Experimental Child Psychology* 116,  
1108 264–277.

- 1109 Lemasson, A., Glas, L., Barbu, S., Lacroix, A., Guilloux, M., Remeuf, K.,
- 1110 Koda, H., 2011. Youngsters do not pay attention to conversational rules:
- 1111 is this so for nonhuman primates? *Nature Scientific Reports* 1.
- 1112 Levelt, W.J., 1989. Speaking: From intention to articulation. MIT press.
- 1113 Levinson, S.C., 2006. On the human “interaction engine”, in: Enfield, N.,
- 1114 Levinson, S. (Eds.), *Roots of Human Sociality: Culture, Cognition and*
- 1115 *Interaction*. Oxford: Berg, pp. 39–69.
- 1116 Levinson, S.C., 2013. Action formation and ascriptions, in: Stivers, T., Sid-
- 1117 nell, J. (Eds.), *The Handbook of Conversation Analysis*. Wiley-Blackwell,
- 1118 Malden, MA, pp. 103–130.
- 1119 Levinson, S.C., 2016. Turn-taking in Human Communication – Origins and
- 1120 Implications for Language Processing. *Trends in Cognitive Sciences* 20,
- 1121 6–14.
- 1122 Magyari, L., Bastiaansen, M.C.M., De Ruiter, J.P., Levinson, S.C., 2014.
- 1123 Early anticipation lies behind the speed of response in conversation. *Jour-*
- 1124 *nal of Cognitive Neuroscience* 26, 2530–2539.

- 1125 Magyari, L., De Ruiter, J.P., 2012. Prediction of turn-ends based on anticipation of upcoming words. *Frontiers in Psychology* 3:376, 1–9.
- 1126
- 1127 Masataka, N., 1993. Effects of contingent and noncontingent maternal stimulation on the vocal behaviour of three-to four-month-old Japanese infants.
- 1128
- 1129 *Journal of Child Language* 20, 303–312.
- 1130
- 1131 Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoni, J., Amiel-Tison, C., 1988. A precursor of language acquisition in young infants.
- 1132 *Cognition* 29, 143–178.
- 1133
- 1134 Morgan, J.L., Saffran, J.R., 1995. Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. *Child Development* 66, 911–936.
- 1135
- 1136 Nazzi, T., Ramus, F., 2003. Perception and acquisition of linguistic rhythm by infants. *Speech Communication* 41, 233–243.
- 1137
- 1138 R Core Team, 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL:
- 1139
- 1140 <http://www.R-project.org>. [Computer program] Version 3.1.1.

- 1141 Ratner, N., Bruner, J., 1978. Games, social exchange and the acquisition of  
1142 language. *Journal of Child Language* 5, 391–401.
- 1143 Ross, H.S., Lollis, S.P., 1987. Communication within infant social games.  
1144 *Developmental Psychology* 23, 241.
- 1145 Rossano, F., Brown, P., Levinson, S.C., 2009. Gaze, questioning and culture,  
1146 in: Sidnell, J. (Ed.), *Conversation Analysis: Comparative Perspectives*.  
1147 Cambridge University Press, Cambridge, pp. 187–249.
- 1148 Sacks, H., Schegloff, E.A., Jefferson, G., 1974. A simplest systematics for the  
1149 organization of turn-taking for conversation. *Language* 50, 696–735.
- 1150 Schegloff, E.A., 2007. *Sequence organization in interaction: Volume 1: A  
1151 primer in conversation analysis*. Cambridge University Press.
- 1152 Shatz, M., 1979. How to do things by asking: Form-function pairings in  
1153 mothers' questions and their relation to children's responses. *Child Develop-  
1154 opment* 50, 1093–1099.
- 1155 Shi, R., Melancon, A., 2010. Syntactic categorization in French-learning  
1156 infants. *Infancy* 15, 517–533.

- 1157 Snow, C.E., 1977. The development of conversation between mothers and  
1158 babies. *Journal of Child Language* 4, 1–22.
- 1159 Soderstrom, M., Seidl, A., Kemler Nelson, D.G., Jusczyk, P.W., 2003. The  
1160 prosodic bootstrapping of phrases: Evidence from prelinguistic infants.  
1161 *Journal of Memory and Language* 49, 249–267.
- 1162 Speer, S.R., Ito, K., 2009. Prosody in first language acquisition—Acquiring  
1163 intonation as a tool to organize information in conversation. *Language and  
1164 Linguistics Compass* 3, 90–110.
- 1165 Stivers, T., Enfield, N.J., Brown, P., Englert, C., Hayashi, M., Heinemann,  
1166 T., Hoymann, G., Rossano, F., De Ruiter, J.P., Yoon, K.E., et al., 2009.  
1167 Universals and cultural variation in turn-taking in conversation. *Proceed-  
1168 ings of the National Academy of Sciences* 106, 10587–10592.
- 1169 Stivers, T., Rossano, F., 2010. Mobilizing response. *Research on Language  
1170 and Social Interaction* 43, 3–31.
- 1171 Takahashi, D.Y., Narayanan, D.Z., Ghazanfar, A.A., 2013. Coupled oscillator  
1172 dynamics of vocal turn-taking in monkeys. *Current Biology* 23, 2162–2168.
- 1173 Thorgrímsson, G., Fawcett, C., Liszkowski, U., 2015. 1- and 2-year-olds'

1174 expectations about third-party communicative actions. *Infant Behavior*  
1175 and Development

1176 Tice (Casillas), M., Henetz, T., 2011. Turn-boundary projection: Looking

1177 ahead, in: *Proceedings of the 33rd Annual Meeting of the Cognitive Science*  
1178 Society, pp. 838–843.

1179 Toda, S., Fogel, A., 1993. Infant response to the still-face situation at 3 and  
1180 6 months. *Developmental Psychology* 29, 532.

1181 Torreira, F., Bögels, S., Levinson, S.C., 2015. Intonational phrasing is neces-  
1182 sary for turn-taking in spoken interaction. *Journal of Phonetics* 52, 46–57.

1183 Weisleder, A., 2012. Richer language experience leads to faster understand-  
1184 ing: Links between language input, processing efficiency, and vocabulary  
1185 growth. Ph.D. thesis. Stanford University.

1186 Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H., 2006.  
1187 Elan: a professional framework for multimodality research, in: *Proceedings*  
1188 of LREC.

1189 **Appendix A. Permutation Analyses**

1190 How can we be sure that our primary dependent measure (anticipatory  
1191 gaze switching) actually relates to turn transitions? Even if children were  
1192 gazing back and forth randomly during the experiment, we would have still  
1193 captured some false hits—switches that ended up in the turn-transition win-  
1194 dows by chance.

1195 We estimated the baseline probability of making an anticipatory switch  
1196 by randomly permuting the placement of the transition windows within each  
1197 stimulus (Figure 4). We then used the switch identification procedure from  
1198 Experiments 1 and 2 (Section 2.1.4) to find out how often participants made  
1199 “anticipatory” switches within these randomly permuted windows. This pro-  
1200 cedure de-links participants’ gaze data from turn structure by randomly re-  
1201 assigning the onset time of each turn-transition in each permutation. We  
1202 created 5,000 of these permutations for each experiment to get an antici-  
1203 patory switch baselines over all possible starting points.

1204 Importantly, the randomized windows were not allowed to overlap with  
1205 each other, keeping true to the original stimuli. We also made sure that the  
1206 properties of each turn transition stayed constant across permutations. So,  
1207 while “transition window A” might start 2 seconds into Random Permuta-

1208 tion 1 and 17 seconds into Random Permutation 2, it maintained the same  
1209 prior speaker identity, transition type, gap duration, language condition, etc.,  
1210 across both permutations.

1211 We then re-ran the statistical models from the original data on each of the  
1212 random permutations, e.g., using Experiment 1's original model to analyze  
1213 the anticipatory switches from each random permutation of the Experiment  
1214 1 looking data. We could then calculate the proportion of random data  
1215  $z$ -values exceeded by the original  $z$ -value for each predictor. We used the  
1216 absolute value of all  $z$ -values to conduct a two-tailed test. If the original  
1217 effect of a predictor exceeded 95% of the random model effects for that same  
1218 predictor, we deemed that predictor's effect to be significantly different from  
1219 the random baseline (i.e.,  $p < .05$ ).

1220 For example, children's "language condition" effect from Experiment 1  
1221 had a  $z$ -value of  $|3.429|$ , which is greater than 99.9% of all  $|z\text{-value}|$  esti-  
1222 mates from Experiment 1's random permutation models (i.e.,  $p = .001$ ). It is  
1223 therefore highly unlikely that the effect of language condition in the original  
1224 model derived from random gaze shifting.

1225 We used this procedure to derive the random-baseline comparison values  
1226 in the main text (above). However, we ran into two issues along the way:

<sub>1227</sub> first, we had to report  $z$ -values rather than beta estimates. Second, we had  
<sub>1228</sub> to exclude a substantial portion of the models, especially in Experiment 2  
<sub>1229</sub> because of model non-convergence. We address each of these issues below.

<sub>1230</sub> *Appendix A.1. Beta, standard error, and  $z$  estimates*

<sub>1231</sub> We reported  $z$ -values in the main text rather than beta estimates because  
<sub>1232</sub> the standard errors in the randomly permuted data models were much higher  
<sub>1233</sub> than for the original data. The distributions for each predictor's beta esti-  
<sub>1234</sub> mate, standard error, and  $z$ -value for adults and children in each experiment  
<sub>1235</sub> are shown in the graphs below (Figures A.1a–A.6b). In each plot, the gray  
<sub>1236</sub> dots represent the absolute value of the 5,000 randomly permuted model es-  
<sub>1237</sub> timates for the estimate type plotted (beta, standard error, or  $z$ ), the white  
<sub>1238</sub> circles represent the model estimates from the original data, and the black  
<sub>1239</sub> triangles represent the 95th percentile for each random distribution.

### Experiment 1: $z$ -value estimates

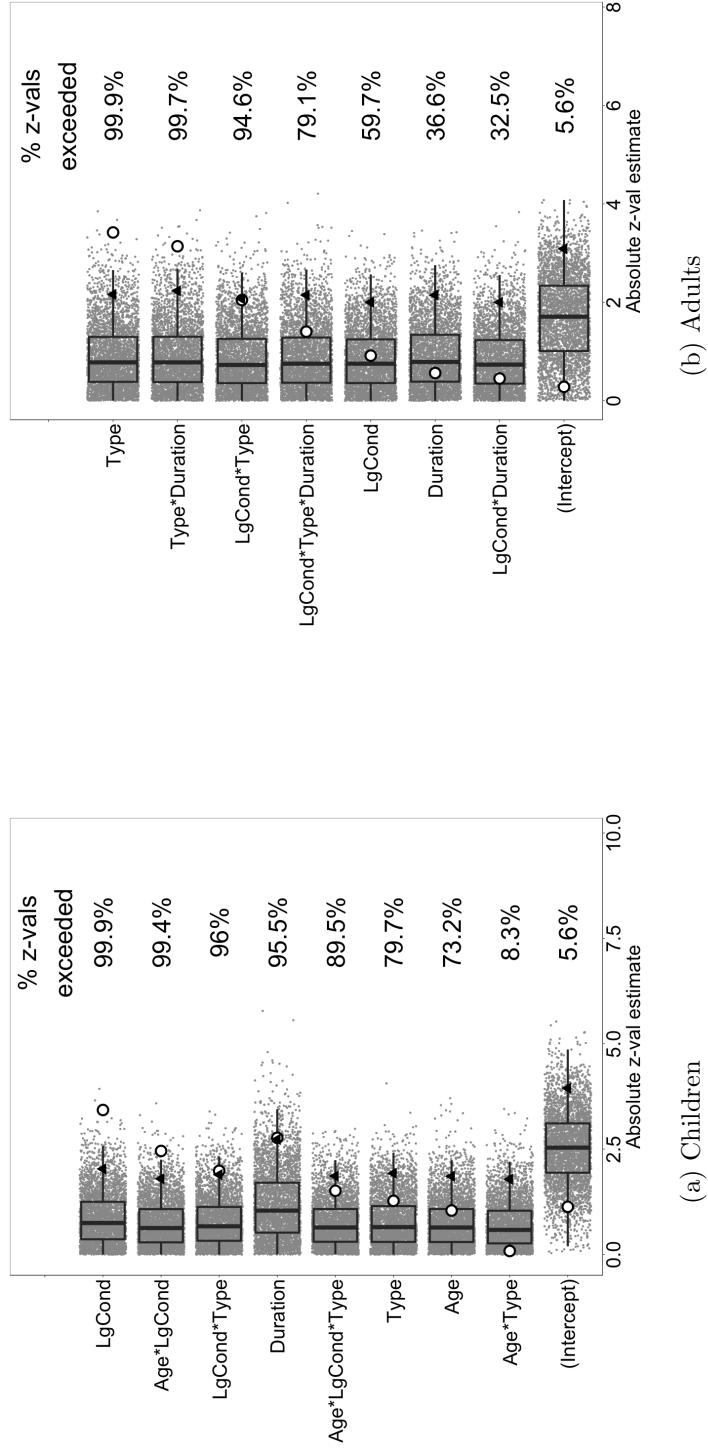


Figure A.1: Random-permutation and original  $|z$ -values for predictors of anticipatory gaze rates in Experiment 1.

### Experiment 1: $\beta$ estimates

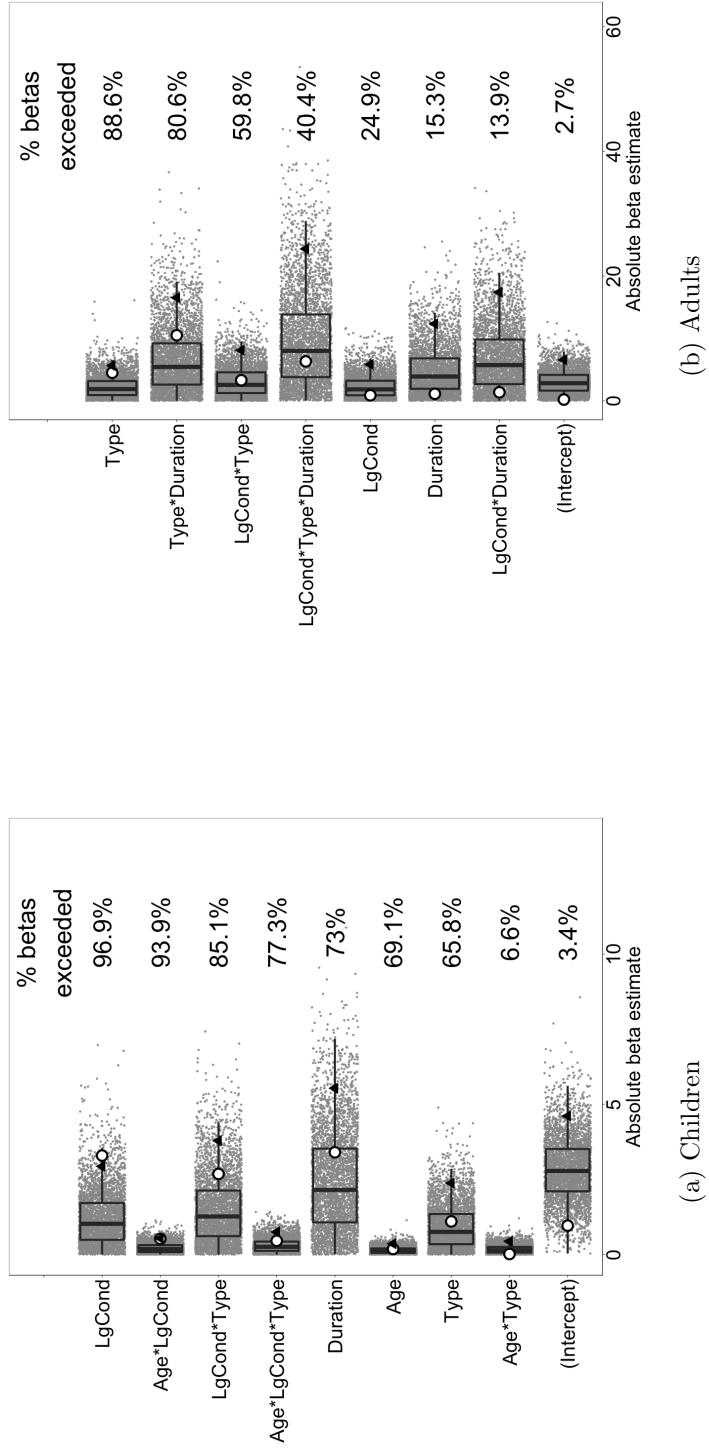


Figure A.2: Random-permutation and original  $|\beta\text{-values}|$  for predictors of gaze rates in Experiment 1.

### Experiment 1: *SE* estimates

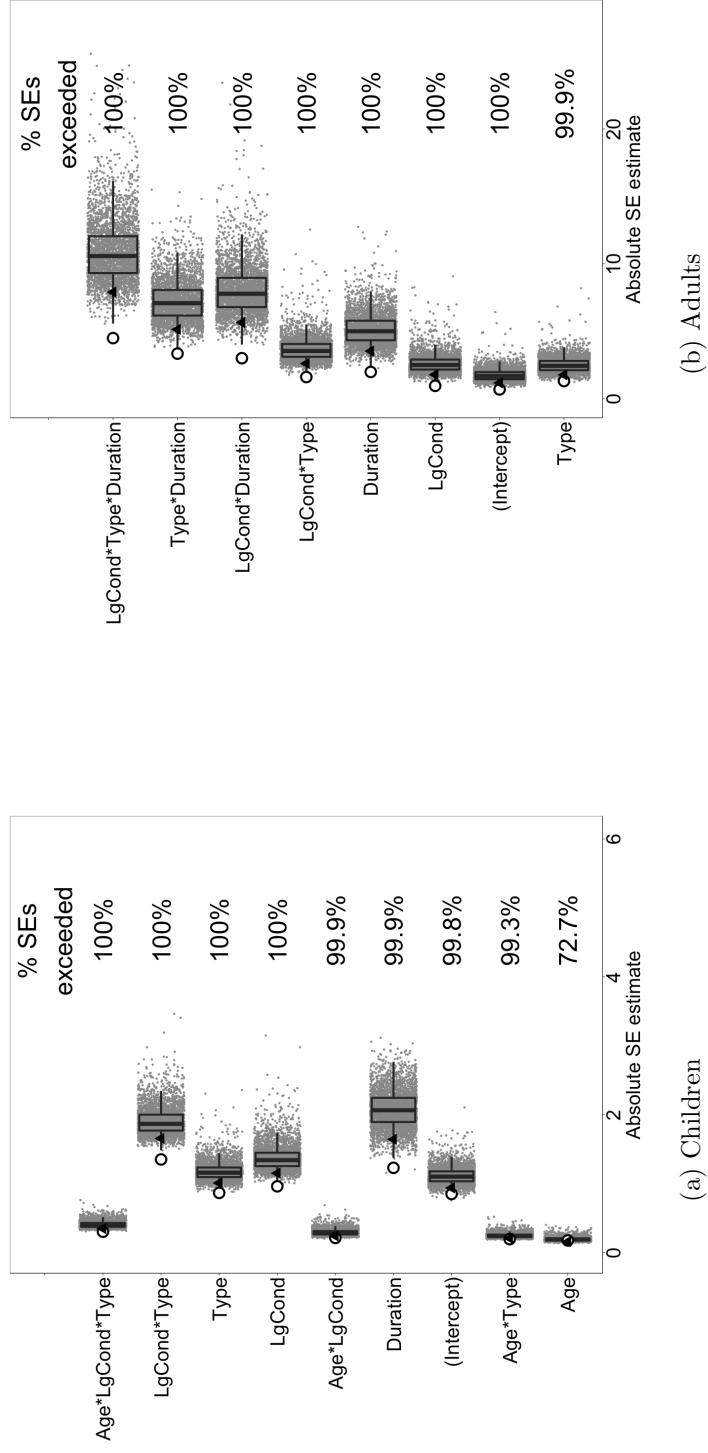


Figure A.3: Random-permutation and original *SE*-values for predictors of anticipatory gaze rates in Experiment 1.

## Experiment 2: $z$ estimates

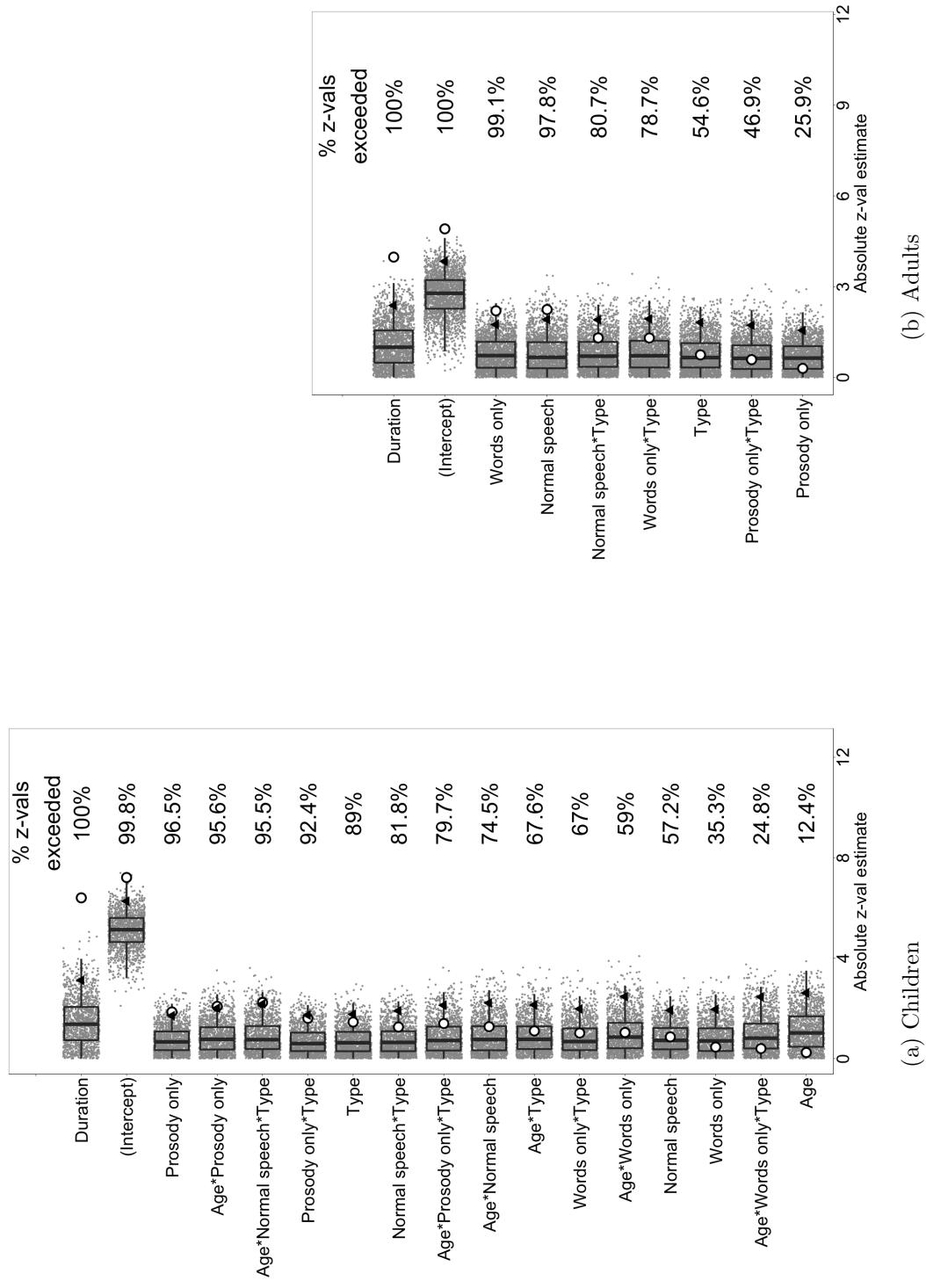


Figure A.4: Random-permutation and original  $|z\text{-values}|$  for predictors of anticipatory gaze rates in Experiment 2.

## Experiment 2: $\beta$ estimates

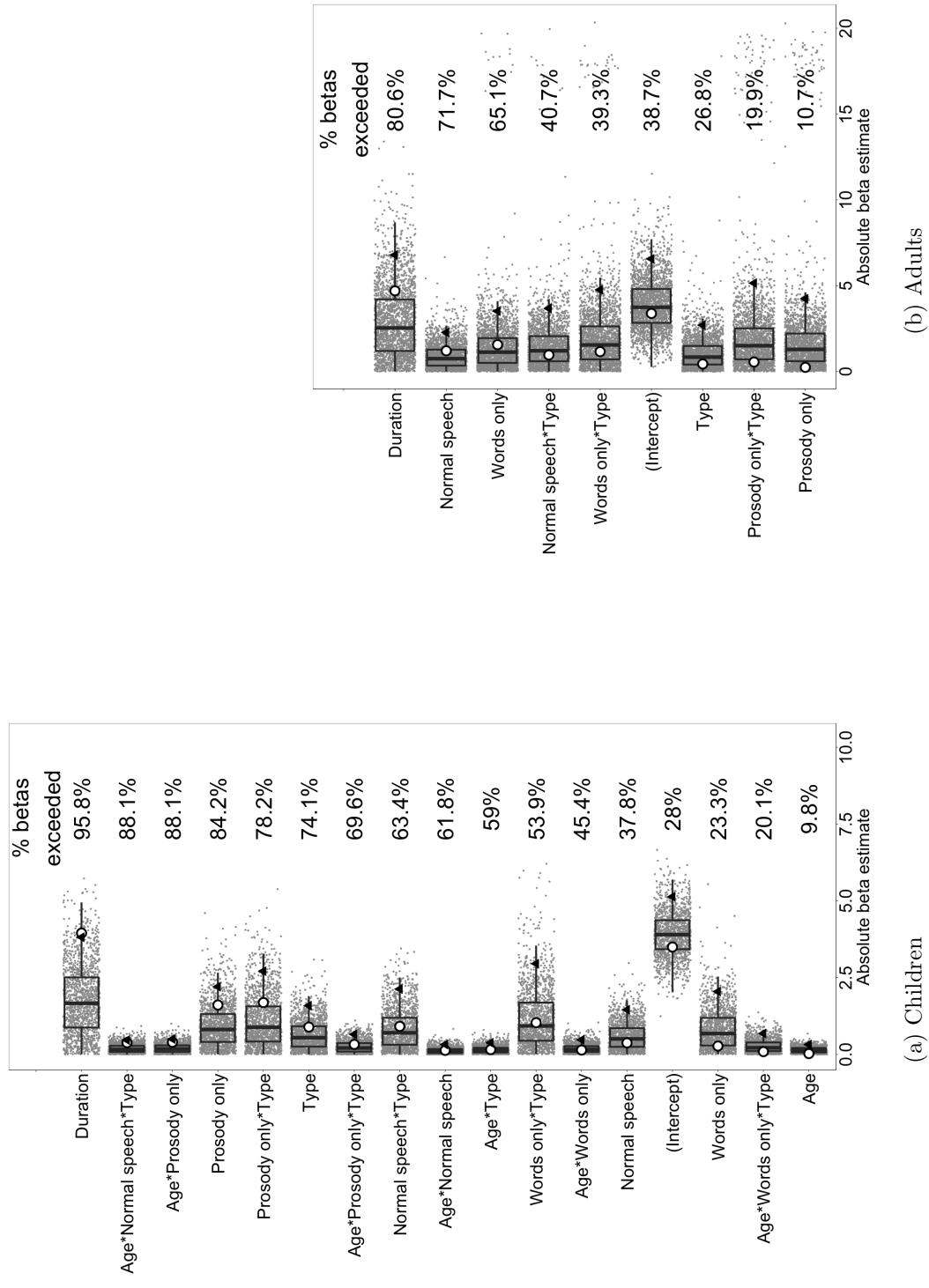


Figure A.5: Random-permutation and original  $|\beta\text{-values}|$  for predictors of anticipatory gaze rates in Experiment 2.

## Experiment 2: *SE* estimates

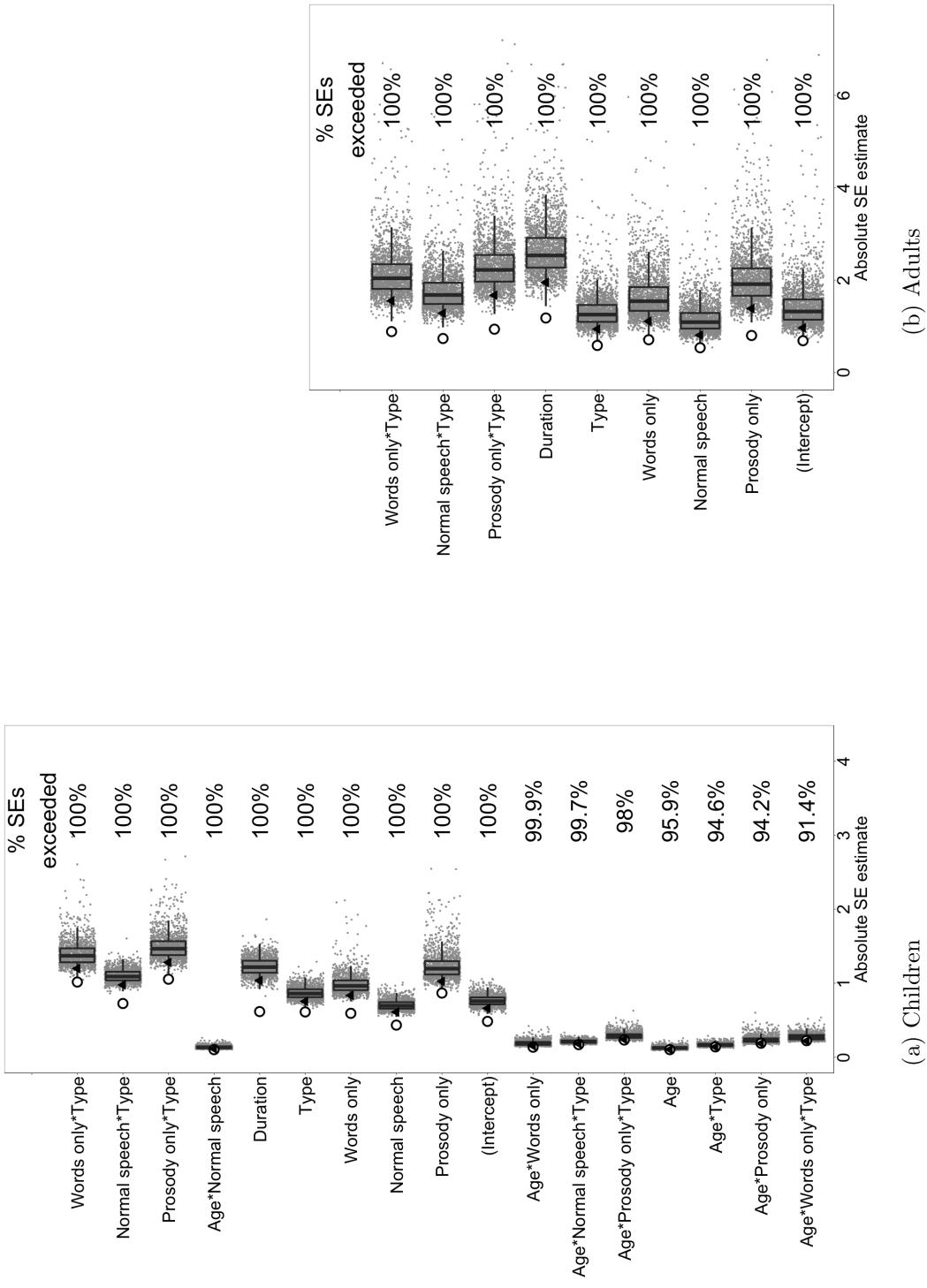


Figure A.6: Random-permutation and original *SE*-values for predictors of anticipatory gaze rates in Experiment 2.

1240 *Appendix A.2. Non-convergent models*

1241 In comparing the real and randomly permuted datasets, we excluded the  
1242 output of random-permutation models that gave convergence warnings to  
1243 remove erratic model estimates from our analyses. Non-convergent models  
1244 made up 22.4–24.4% of the random permutation models in Experiment 1 and  
1245 69–70% of the random permutation models in Experiment 2. The  $z$ -values for  
1246 each predictor in the converging and non-converging models from Experiment  
1247 1 are shown in Table A.1.

1248 Although many of the non-converging models show estimates within range  
1249 of the converging models (e.g., with a mean difference of only 0.09 in median  
1250  $z$ -value across predictors), they also show many radically outlying estimates  
1251 (e.g., showing a mean difference of 237.3 in mean  $z$ -value across predictors).  
1252 Similar patterns were obtained in the non-converging models for Experiment  
1253 2 and persisted even when we tried other optimizers.

1254 We suspect that the issue derives from data sparsity in some of the ran-  
1255 dom permutations. This problem is known to occur when there are limited  
1256 numbers of binary observations in each of a design matrix's bins (Allison,  
1257 2004). We could instead use zero-inflated poisson or negative binomial re-  
1258 gression models to allow for overdispersion in our data (Allison, 2012). How-

<sub>1259</sub> ever, these would give us baselines for the normal, convergent model, which

<sub>1260</sub> is not the aim of this analysis.

	Mean <sub>C</sub>	Mean <sub>NC</sub>	Median <sub>C</sub>	Median <sub>NC</sub>	SD <sub>C</sub>	SD <sub>NC</sub>	Min <sub>C</sub>	Min <sub>NC</sub>	Max <sub>C</sub>	Max <sub>NC</sub>
<i>Children</i>										
(Intercept)	-2.52	-458.42	-2.54	-2.86	0.87	1319.22	-5.53	-8185.36	0.41	0.97
Age	-0.51	-17.83	-0.49	-0.53	0.79	83.78	-3.71	-672.2	2.3	342.8
LgCond	-0.53	-109.91	-0.55	-0.63	0.93	564.42	-3.93	-4418.74	3.23	2296.19
Type	-0.1	-29.66	-0.09	-0.1	0.98	515.12	-4.06	-4383.92	3.36	3416.68
Duration	0.99	345.53	0.98	1.15	1.07	1323.13	-2.44	-5048.24	5.78	9985.16
Age*LgCond	0.19	10.64	0.2	0.18	0.9	109.6	-3.31	-581.61	3.59	946.81
Age*Type	0.02	-1.8	0.001	-0.04	0.9	98.27	-3.36	-884.36	3.45	640.43
LgCond*Type	0.2	45.32	0.2	0.27	0.96	691.3	-3.12	-4160.06	3.39	5107.64
Age*LgCond*Type	-0.12	-14.23	-0.12	-0.15	0.93	156.72	-2.98	-1318.26	2.90	927.69
<i>Adults</i>										
(Intercept)	-1.63	-126.14	-1.71	-1.73	0.97	713.39	-4.08	-12111.22	2.15	649.55
LgCond	-0.26	-679.6	-0.3	-0.53	1.02	15894.33	-3.45	-494979.7	3.35	88581.58
Type	-0.11	6.29	-0.13	-0.04	1.11	501.5	-3.85	-6420.75	3.28	8177.88
Duration	0.25	84.09	0.27	0.26	1.1	1152.94	-3.25	-10864.51	3.46	18540.62
LgCond*Type	0.12	-242.27	0.1	0.34	1.07	26836.7	-3.41	-622642.7	3.81	509198.4
LgCond*Duration	0.15	780.03	0.16	0.39	1.04	44105.02	-3.84	-798498.6	3.55	1145951
Type*Duration	0.05	-6.56	0.05	0.02	1.13	1389.9	-3.54	-15979.22	3.87	16419.46
LgCond*Type*Duration	-0.06	1083.63	-0.08	-0.21	1.1	63116.54	-4.21	-1201895	4.02	1284965

Table A.1: Estimated  $z$ -values for each predictor in converging ( $C$ ) and non-converging ( $NC$ ) child and adult models from Experiment 1.

1261 **Appendix B. Pairwise developmental tests**

1262 Experiments 1 and 2 both showed effects of age in interaction with lin-  
1263 guistic condition and transition type (e.g., English vs. non-English). To  
1264 explore these effects in more depth, we recorded the average difference score  
1265 for the predictor that interacted with age for each participant (e.g., English  
1266 minus non-English anticipatory switches), using these values to compute an  
1267 average difference score over participants in each age group (e.g., age 3, 4,  
1268 and 5) within each random permutation. That averaging process produces  
1269 5,000 baseline-derived difference scores for each age group.

1270 We then made pairwise age comparisons of these difference scores (e.g.,  
1271 the linguistic condition effect in 3-year-olds vs. 4-year-olds), computing the  
1272 percent of random-permutation difference scores exceeded by the real-data  
1273 difference score. If the real-data difference score exceeded 95% of the random-  
1274 data age difference scores, we deemed it to be an age effect significantly dif-  
1275 ferent from chance—e.g., a significant difference between ages three and four  
1276 in the effect of linguistic condition. This procedure is essentially a two-tailed  
1277 *t*-test, adapted for use with the randomly permuted baseline data.

1278 In each of the plots below, the black dot represents the real data value  
1279 for the effect being shown. The effect sizes from the 5,000 randomly per-

<sub>1280</sub> muted data sets are shown in the distribution. The percentage displayed  
<sub>1281</sub> is the percentage of random permutation values exceeded by the original  
<sub>1282</sub> data value (taking the absolute value of all data points for a two-tailed test).  
<sub>1283</sub> Comparisons marked with 95% or higher are significant at the  $p<0.05$  level.

Experiment 1: Age and linguistic condition

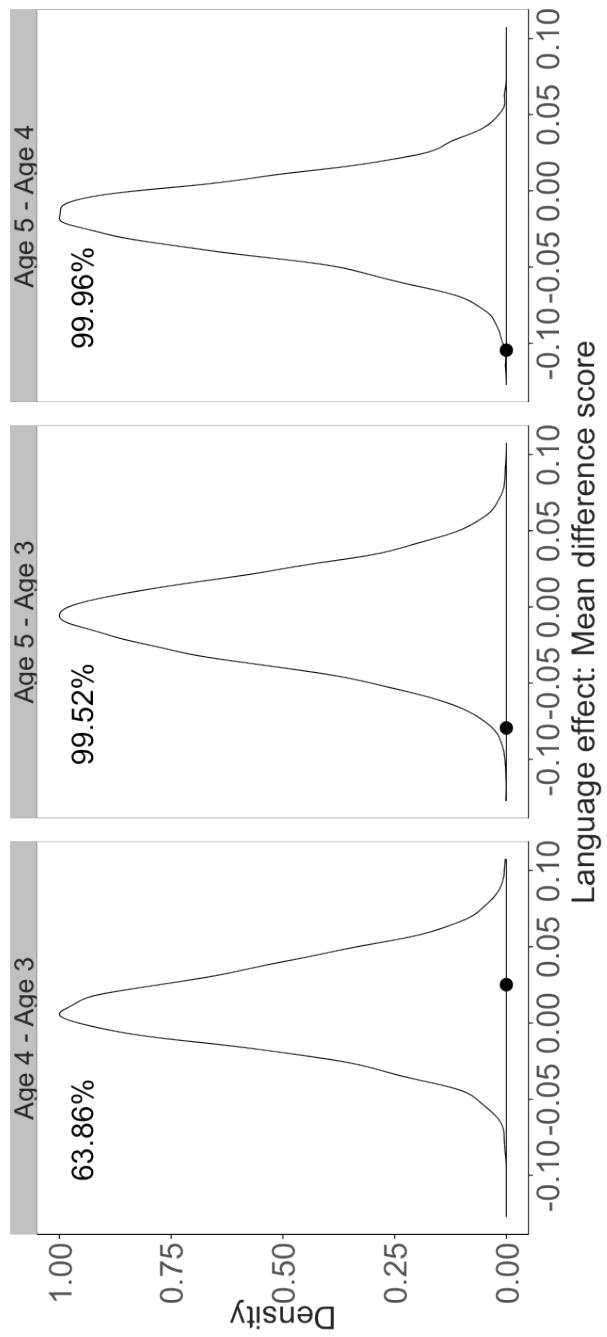


Figure B.1: Pairwise comparisons of the language condition effect across ages in Experiment 1.

### Experiment 2: Age and the prosody only condition

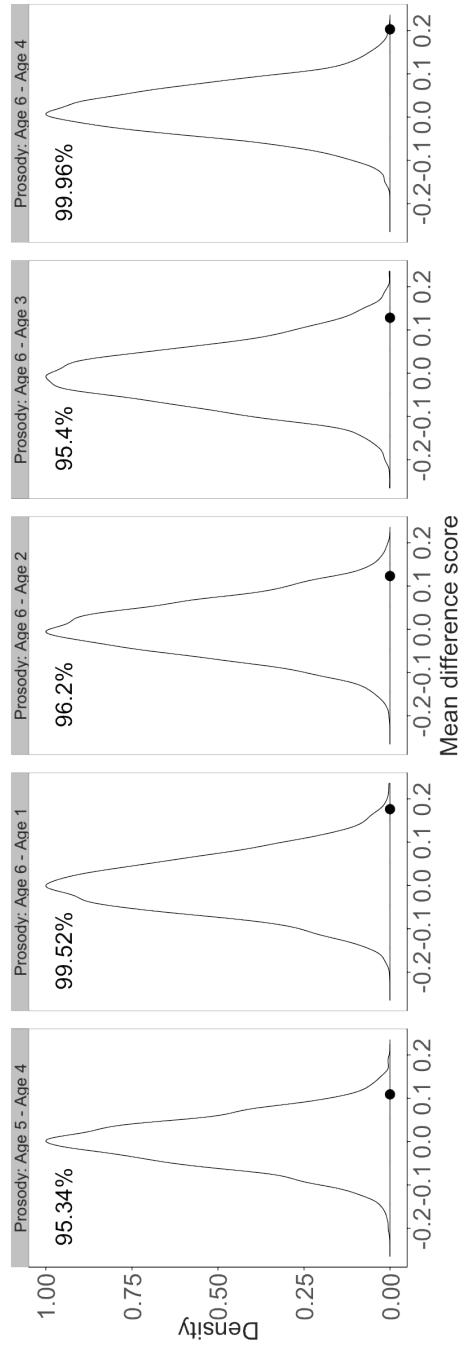


Figure B.2: Significant pairwise comparisons of the *prosody only-no speech* linguistic condition effect, across ages in Experiment 2

## Experiment 2: Age, transition type, and *normal* speech

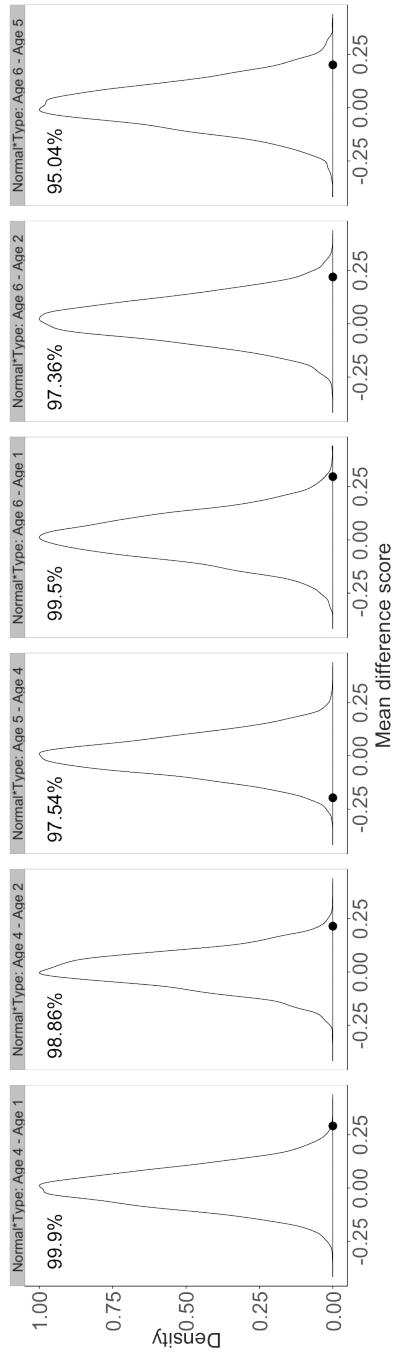


Figure B.3: Significant pairwise comparisons of the *normal speech-no speech* language condition effect for transition type, across ages, in Experiment 2.

1284 **Appendix C. Boredom-driven anticipatory looking**

1285 One alternative hypothesis for children's anticipatory gazes is that they  
1286 simply grow bored and start looking away at a constant rate after a turn  
1287 begins. This data plotted here show a hypothetical group of boredom-driven  
1288 participants (gray dots) compared to participants from the actual data in  
1289 Experiment 2 (black dots). The hypothetical boredom-driven participants  
1290 look away from the current speaker at a linear rate, beginning one second  
1291 after the start of a turn.

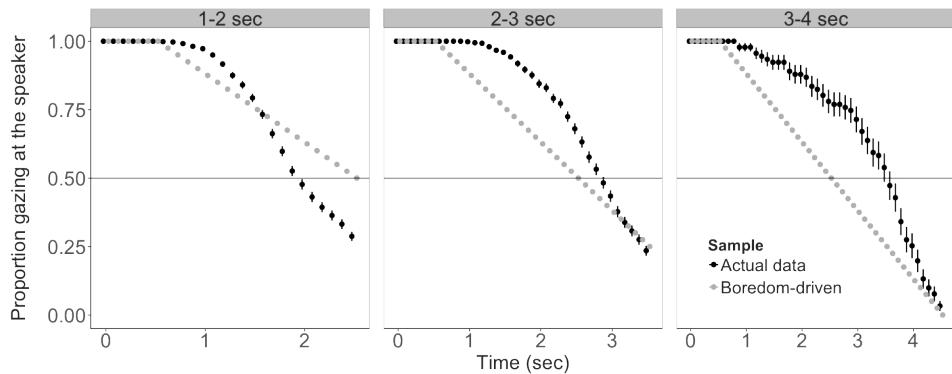


Figure C.1: Proportion of participants (hypothetical boredom-driven=gray; actual Experiment 2=black) looking at the current speaker, split by turn duration. Vertical bars indicate standard error in the experimental data.

1292 If children's switches away from the current speaker were driven by bore-

1293 dom, they would switch away equally quickly on long and short turns. How-  
1294 ever, as turn duration increased, children in the experiment looked at the  
1295 current speaker for a longer period before looking away, suggesting that they  
1296 were not switching away at a constant rate. We can see this pattern most  
1297 clearly in the time at which 50% of the children had switched away from  
1298 the current (indicated by the horizontal line in Figure C.1): children's gaze  
1299 crossed this 50% threshold at 2.0, 2.9, and 3.6 seconds after the start of speech  
1300 for turns with durations of 1–2, 2–3, and 3–4 seconds, respectively. In con-  
1301 trast, the hypothetical boredom-driven children always hit the 50% threshold  
1302 at 2.5 seconds after the start of speech. This pattern suggests that, though  
1303 children do look away with time, their looks away are not simply driven by  
1304 boredom.

1305 **Appendix D. Puppet pair and linguistic condition**

1306 The design for Experiment 2 does not fully cross puppet pair (e.g., robots,  
1307 blue puppets) with linguistic condition (e.g., *words only* and *no speech*). Even  
1308 though each puppet pair is associated with different conversation clips across  
1309 children (e.g., robots talking about kitties, birthday parties, and pancakes),  
1310 the robot puppets themselves were exclusively associated with the *words only*  
1311 condition. Similarly, merpeople were exclusively associated with *prosody only*  
1312 speech, and the puppets wearing dressy clothes were exclusively associated  
1313 with the *no speech* condition. We designed the experiment this way to in-  
1314 crease its pragmatic felicity for older children (i.e., robots make robot sounds,  
1315 merpeople’s voices are muffled under the water, the party-going puppets are  
1316 in a ‘party’ room with many other voices). There is therefore a confound  
1317 between linguistic condition and puppet pair; for example, children could  
1318 have made fewer anticipatory switches in the *prosody only* condition because  
1319 the puppets were less interesting. To test whether puppet pair drove the  
1320 condition-based differences found in Experiment 2, we ran a short follow-up  
1321 study.

1322 **Methods**

1323 We recruited 30 children between ages 3;0 and 5;11 from the Children's Dis-  
1324 covery Museum of San Jose, California to participate in our experiment. All  
1325 participants were native English speakers. Children were randomly assigned  
1326 to one of six videos (five children per video).

1327 *Materials.* We created 6 short videos from the stimulus recordings made for  
1328 Experiment 2. Each video featured a puppet pair (red/blue/yellow/robot/  
1329 merpeople/party-goer; Figure 5). Puppets in all six videos performed the  
1330 exact same conversation recording ('birthday party'; Experiment 2) with  
1331 normal, unmanipulated speech; this experiment therefore holds all things  
1332 constant across stimuli except for the appearance of the puppets.

1333 *Procedure.* We used the same experimental apparatus and procedure as in  
1334 Experiments 1 and 2. Each participant was randomly assigned to watch only  
1335 one of the six puppet videos. Five children watched each video. As in Exper-  
1336 iment 2, the experimenter immediately began each session with calibration  
1337 and then stimulus presentation because no special instructions were required.  
1338 The entire experiment took less than three minutes.

1339 *Data preparation.* We identified anticipatory gaze switches to the upcoming  
1340 speaker using the same method as in Experiments 1 and 2.

1341 **Results and discussion**

1342 We modeled children’s anticipatory switches (yes or no at each transition)  
1343 with mixed effects logistic regression, including puppet pair (robots/mer-  
1344 people/party-goers/other-3) as a fixed effect and participant and turn tran-  
1345 sition as random effects. We grouped the red, blue, and yellow puppets  
1346 together because they collectively represented the puppets used in the *nor-*  
1347 *mal* speech condition—this follow-up experiment is meant to test whether  
1348 the condition-based differences from Experiment 2 arose from the puppets  
1349 used in each condition.

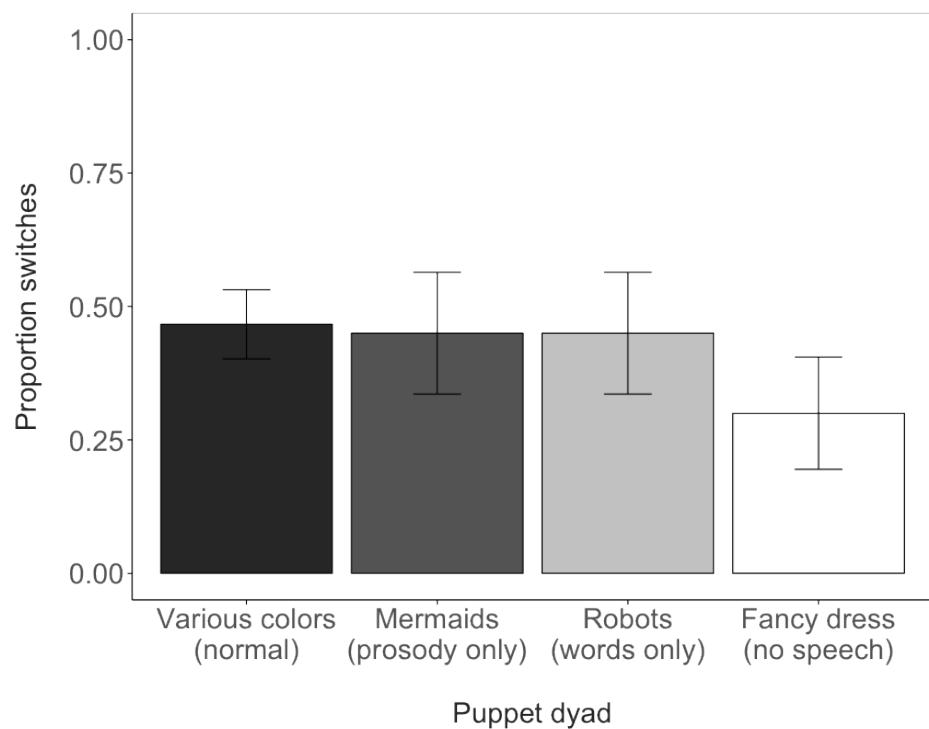


Figure D.1: Proportion gaze switches across puppet pairs when linguistic condition and conversation are held constant.

	Estimate	Std. Error	<i>z</i> value	Pr(>  <i>z</i>  )
<i>Reference level: normal-condition puppets</i>				
(Intercept)	-0.14790	0.32796	-0.451	0.652
Puppets= <i>mermaid</i>	-0.07581	0.65532	-0.116	0.908
Puppets= <i>robot</i>	-0.07104	0.65321	-0.109	0.913
Puppets= <i>party</i>	-0.78206	0.68699	-1.138	0.255
<i>Reference level: mer-puppets</i>				
(Intercept)	-0.22371	0.56832	-0.394	0.694
Puppets= <i>robot</i>	0.004763	0.80096	0.006	0.995
Puppets= <i>party</i>	-0.70626	0.82742	-0.854	0.393
<i>Reference level: robot puppets</i>				
(Intercept)	-0.21895	0.56565	-0.387	0.699
Puppets= <i>party</i>	-0.71102	0.82657	-0.860	0.390
<i>Reference level: party-goer puppets</i>				
(Intercept)	-0.9300	0.6067	-1.533	0.125

Table D.2: Model output for children’s anticipatory gaze switches with reference levels varied to show all possible pairwise differences between puppet pairs.

1350 In four versions of this model, we systematically varied the reference level  
 1351 of the puppet pair to check for any cross-condition differences. We found no  
 1352 significant effects of puppet pair on switching rate (all  $p > 0.25$ ; Table D.2).  
 1353 We take this finding as evidence that our decision to not fully cross puppet  
 1354 pairs and linguistic conditions in Experiment 2 was unlikely to have strongly  
 1355 affected children’s anticipatory gaze rates above and beyond the intended  
 1356 effects of linguistic condition.