Supplementary Materials

Material A. Permutation Analyses

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

We estimated the baseline probability of making an anticipatory switch by randomly permuting the placement of the transition windows within each stimulus (Main paper; Figure 4). We then used the switch identification procedure from Experiments 1 and 2 to find out how often participants made "anticipatory" switches within these randomly permuted windows. This procedure de-links participants' gaze data from turn structure by randomly reassigning the onset time of each turn-transition in each permutation. We created 5,000 of these permutations for each experiment to get an anticipatory switch baselines over all possible starting points.

Importantly, the randomized windows were not allowed to overlap with each other, keeping true to the original stimuli. We also made sure that the properties of each turn transition stayed constant across permutations. So, while "transition window A" might start 2 seconds into Random Permutation 1 and 17 seconds into Random Permutation 2, it maintained the same prior speaker identity, transition type, gap duration, language condition, etc., across both permutations.

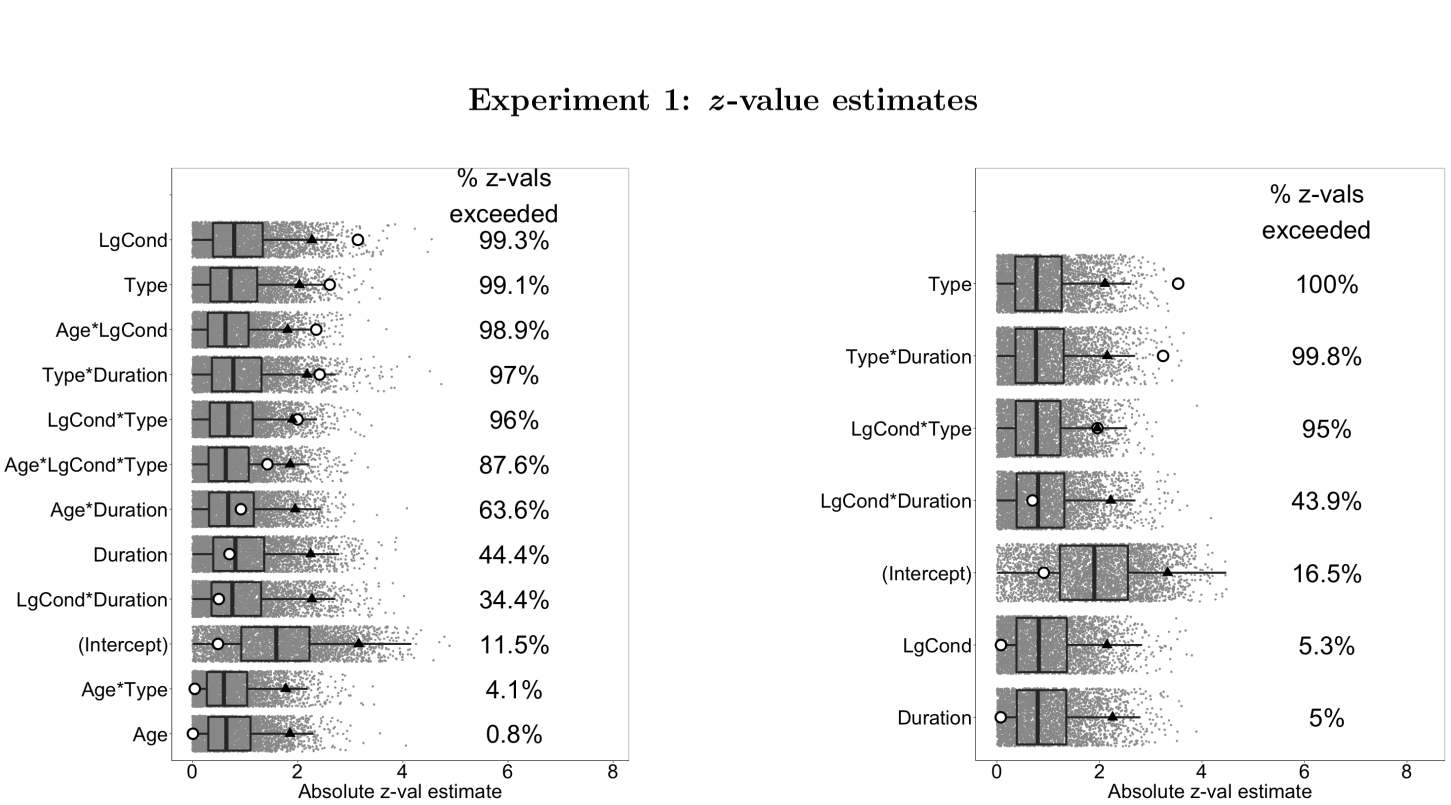
We then re-ran the statistical models from the original data on each of the random permutations, e.g., using Experiment 1's original model structure to analyze the anticipatory switches from each random permutation of the Experiment 1 looking data. We could then calculate the proportion of random data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

For example, children's "language condition" effect from Experiment 1 had a z-value of $|3.65|$, which is greater than 99.3% of all $|z\text{-value}|$ estimates from Experiment 1's random permutation models (i.e., p=.007). It is therefore highly unlikely that the effect of language condition in the original model came from random gaze shifting.

We used this procedure to derive the random-baseline comparison values in the main text (above). However, we ran into two issues along the way: first, we had to report z-values rather than beta estimates of each effect. Second, we had to exclude a substantial portion of the models, especially in Experiment 2 because of model non-convergence. We address each of these issues below.

Material A.1. Beta, standard error, and z estimates

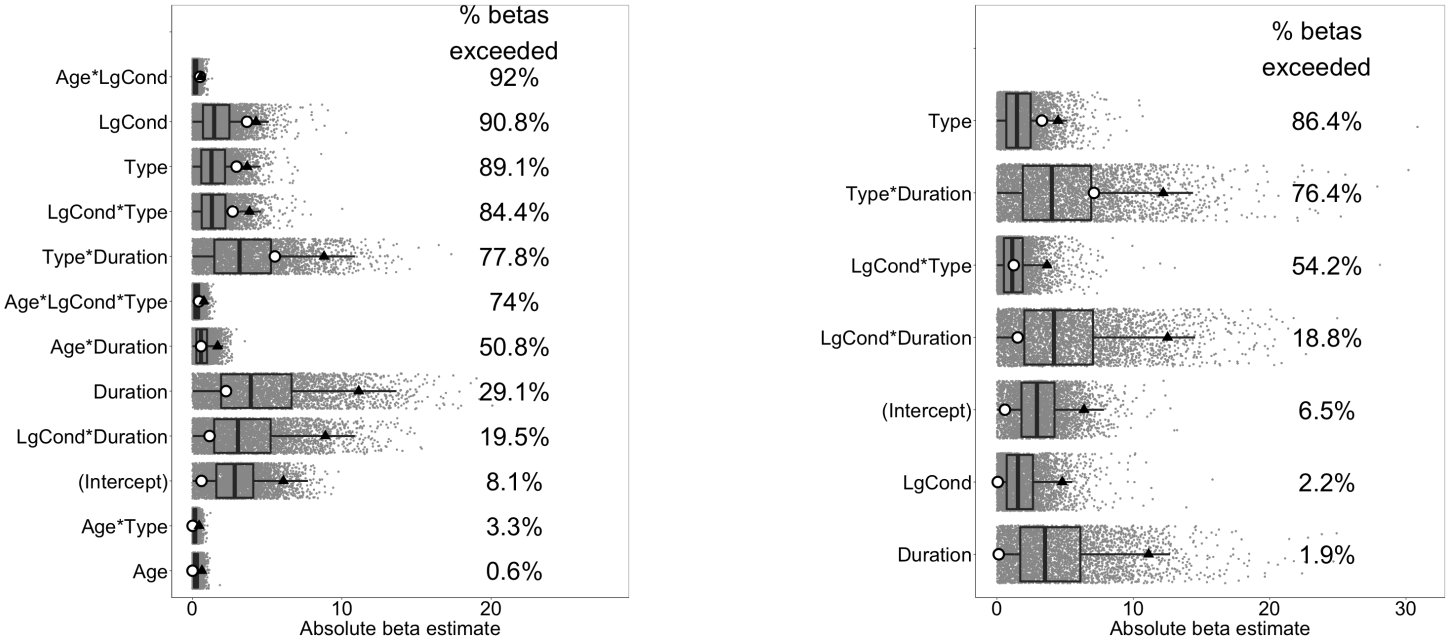
We reported z-values in the main text rather than beta estimates because the standard errors in the randomly permuted data models were much higher than for the original data. The distributions for each predictor's beta estimate, standard error, and z-value for adults and children in each experiment are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



(a) Children (b) Adults

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

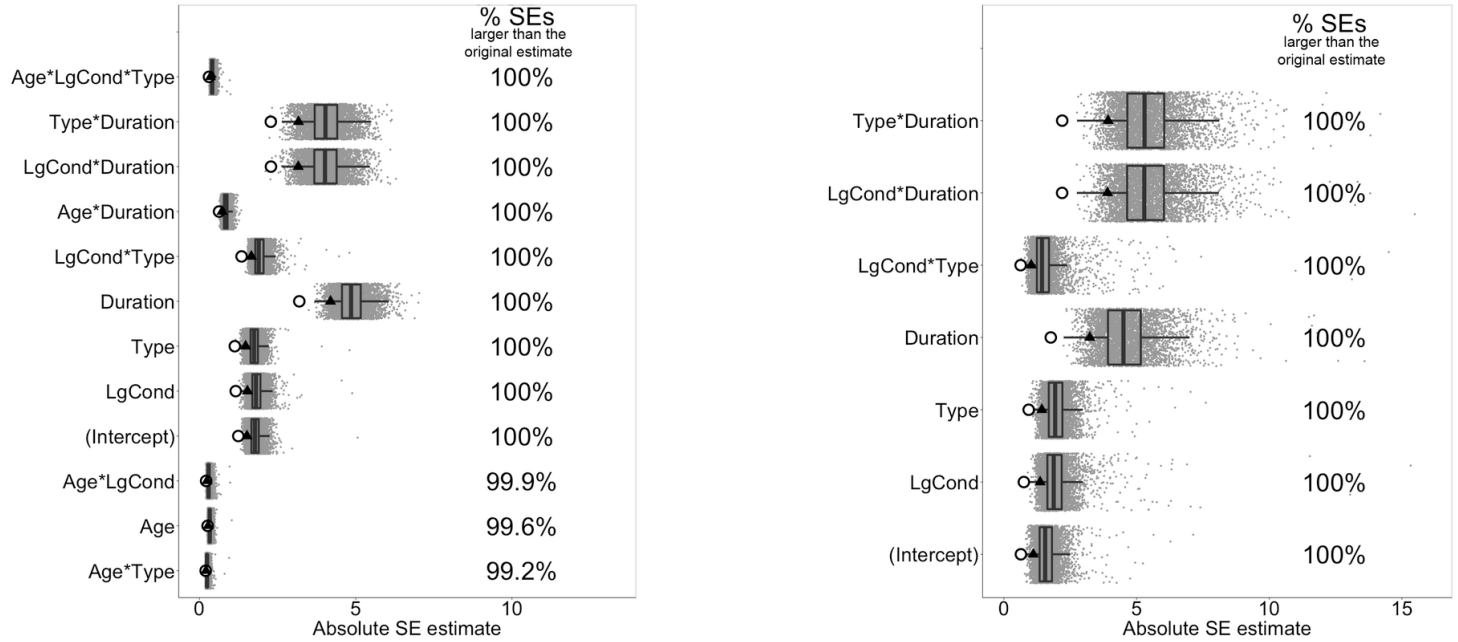
Experiment 1: β estimates



(a) Children (b) Adults

Figure A.2: Random-permutation and original |β-values| for predictors of gaze rates in Experiment 1.

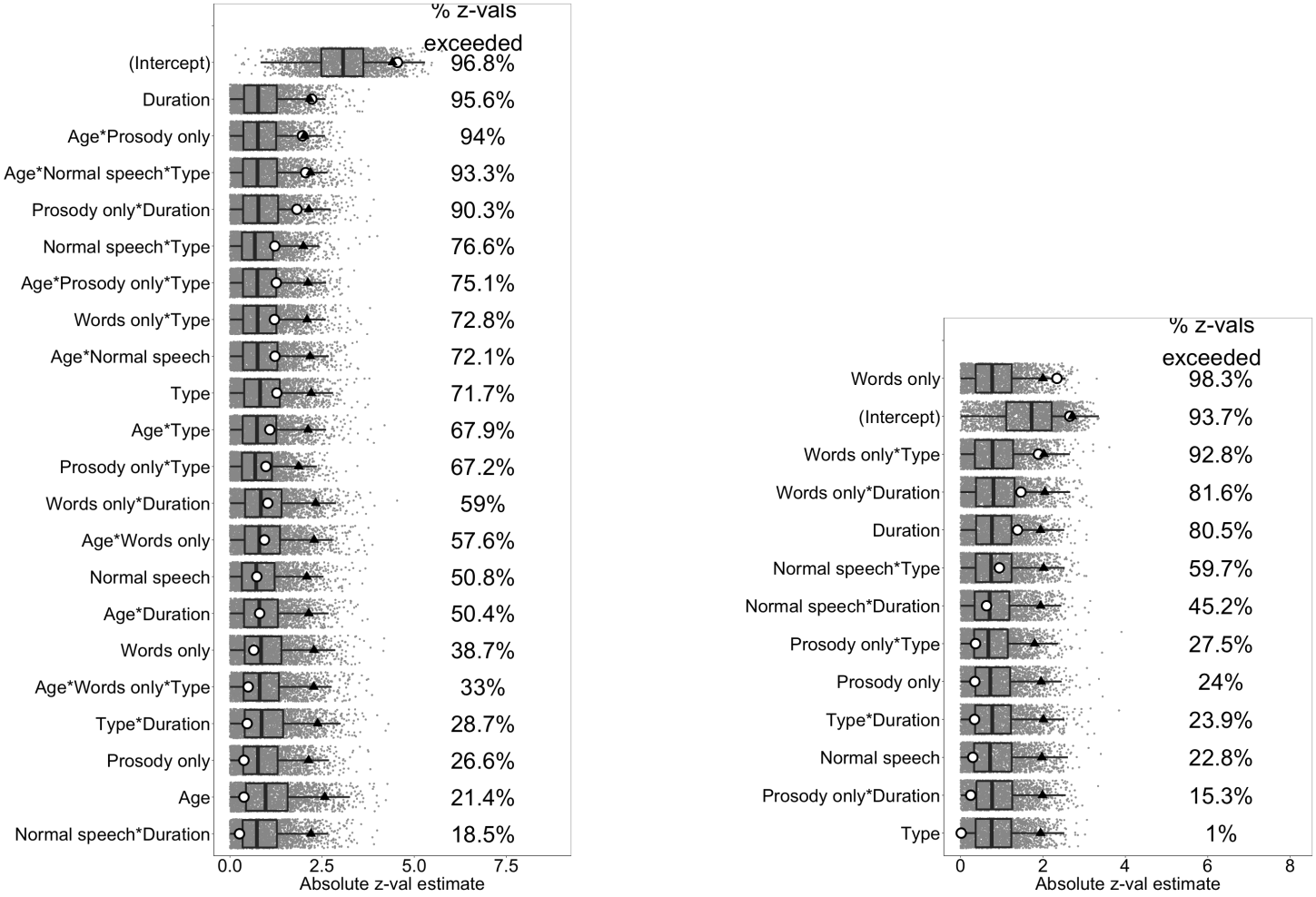
Experiment 1: *SE* estimates



(a) Children (b) Adults

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

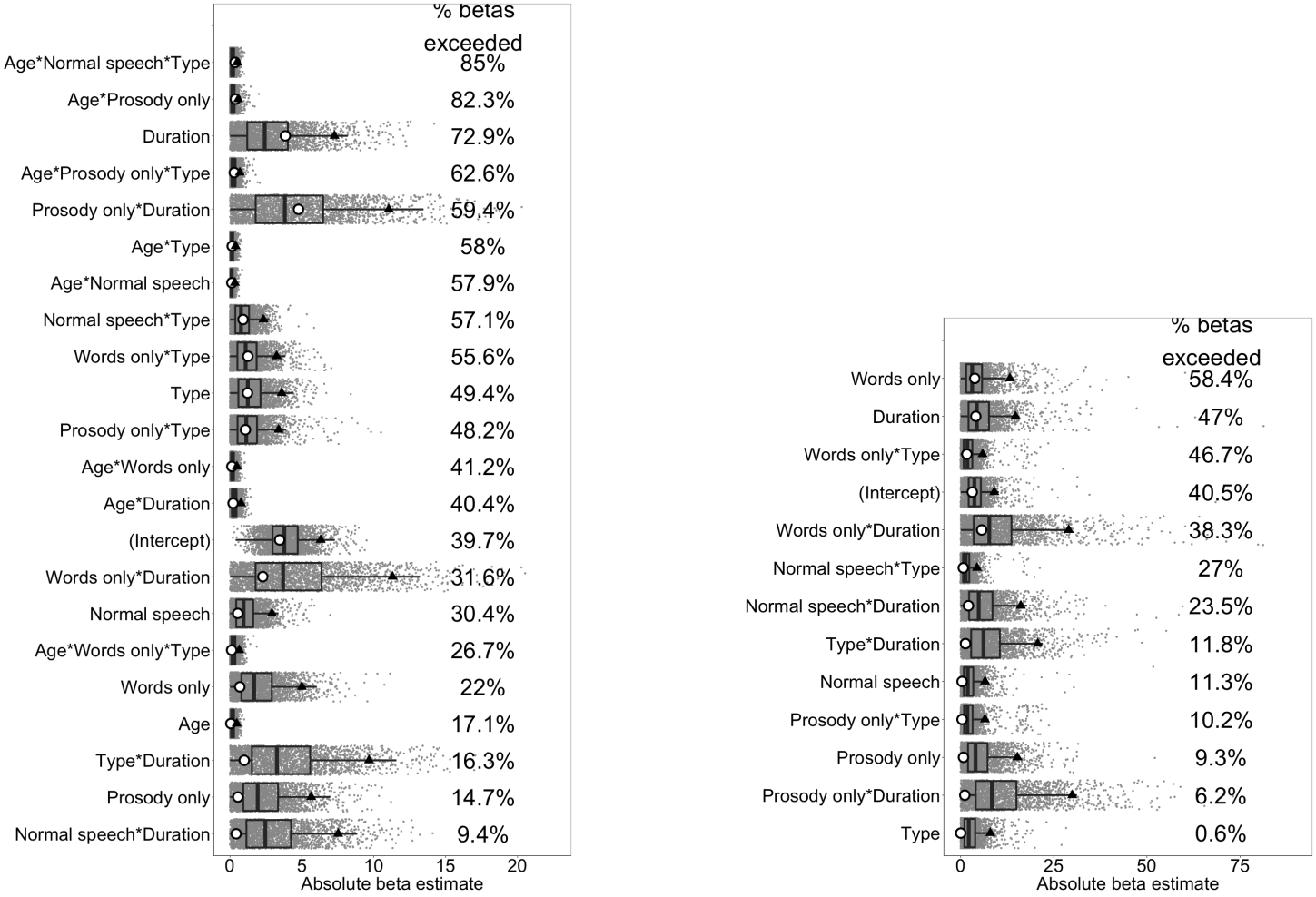
Experiment 2: z estimates



(a) Children (b) Adults

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

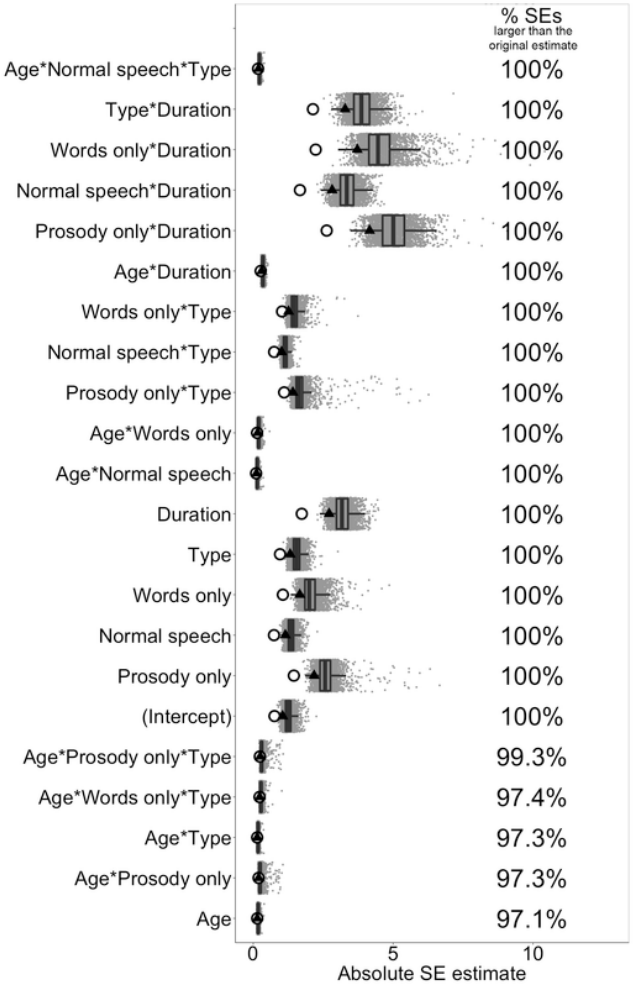
Experiment 2: β estimates



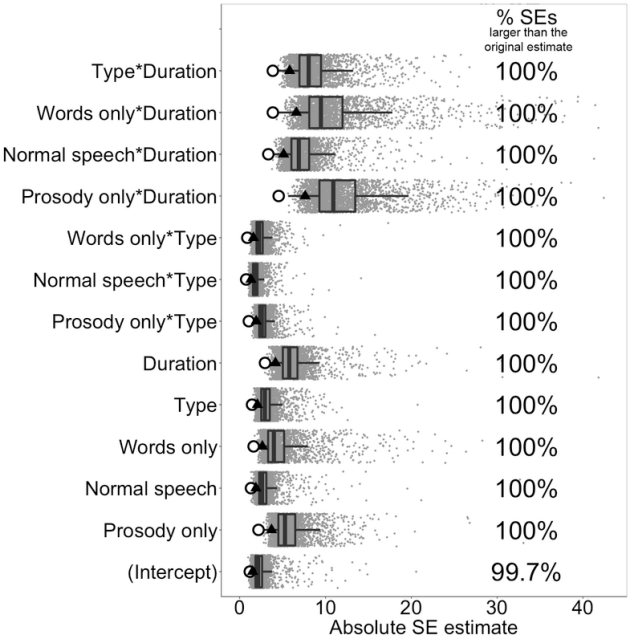
(a) Children (b) Adults

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

Experiment 2: *SE* estimates



(a) Children



(b) Adults

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

Material A.2. Non-convergent models

In comparing the real and randomly permuted datasets, we excluded the output of random-permutation models that gave convergence warnings to remove erratic model estimates from our analyses. Non-convergent models made up 13–14% of the random permutation models in Experiment 1 and 46–49% of the random permutation models in Experiment 2. The z-values for each predictor in the converging and non-converging models from Experiment 1 are shown in Table A.1.

Although many of the non-converging models show estimates within range of the converging models (e.g., with a mean difference of only 0.096 in median z-value across predictors), they also show many radically outlying estimates (e.g., showing a mean difference of 146.7 in mean z-value across predictors). Similar patterns were obtained in the non-converging models for Experiment 2 and persisted across multiple attempts with different optimizers.

We suspect that the issue derives from data sparsity in some of the random permutations. This problem is known to occur when there are limited numbers of binary observations in each of a design matrix's bins (Allison, 2004). We could instead use zero-inflated poisson or negative binomial regression models to allow for overdispersion in our data (Allison, 2012). However, these would give us baselines for the normal, convergent model, which is not the aim of this analysis.

	Mean_C	Mean_{NC}	Median_C	Median_{NC}	SD_C	SD_{NC}	Min_C	Min_{NC}	Max_C	Max_{NC}
Children										
(Intercept)	-1.56	-901.93	-1.59	-1.94	0.98	1945.68	-4.89	-11942.58	2.16	1840.2
Age	-0.2	-26.08	-0.21	-0.28	0.92	193.6	-4.06	-1151.44	3.57	751.7
LgCond	-0.54	-313.61	-0.56	-0.77	1.04	1281.84	-4.55	-7781.18	3.51	4341.3
TType	-0.03	-22.27	-0.03	0.01	1.05	1099.5	-3.42	-7137.95	3.56	5034.84
GapDur	0.42	511.04	0.45	0.61	1.09	3555.2	-3.86	-15899.54	3.88	21151.4
Age*LgCond	0.18	6.46	0.18	0.23	0.91	160.59	-3.35	-791.57	3.69	950.17
Age*TType	0.02	-7.08	-0.01	-0.05	0.9	152.2	-3.45	-815.06	3.43	741.38
LgCond*TType	0.18	-5.76	0.2	0.21	0.97	1129.35	-3.26	-6230.78	3.4	5997.59
Age*GapDur	-0.11	-24.39	-0.08	-0.12	0.99	536.89	-4.08	-2897.34	2.87	2602.11
LgCond*GapDur	0.22	475.09	0.2	0.4	1.12	2988.88	-3.83	-14231.85	4.02	17307.34
Ttype*GapDur	-0.02	-37.07	-0.03	-0.12	1.13	2824.93	-4.51	-16493.61	4.73	14994.45
Age*LgCond*TType	-0.1	-2.92	-0.11	-0.21	0.93	241.44	-3.34	-1434.96	3.02	1333.34
Adults										
(Intercept)	-1.85	-135.7	-1.9	-1.96	0.96	707.63	-4.48	-8056.34	1.61	654.56
LgCond	-0.35	-57.44	-0.37	-0.5	1.09	625.12	-3.8	-6033.9	3.68	5343.37
TType	-0.06	9.59	-0.06	0	1.09	403.93	-3.54	-4131.97	3.34	3793.07
GapDur	0.31	97.73	0.32	0.38	1.12	1159.99	-3.11	-7149.74	3.89	10669.09
LgCond*TType	0.18	31.6	0.18	0.22	1.03	560.99	-2.87	-7722.35	3.9	4377.92
LgCond*GapDur	0.19	77.34	0.21	0.18	1.12	1047.37	-4.18	-7713.96	3.71	7764.19
Ttype*GapDur	0	-50.12	0.01	-0.07	1.11	1065.37	-3.42	-10640.42	3.64	7868.74

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

Material B. Pairwise developmental tests

Experiments 1 and 2 both showed effects of age in interaction with linguistic condition and transition type. To explore these effects in more depth, in each permutation we recorded the average difference score for each participant, for each predictor that interacted with age (e.g., English minus non-English anticipatory switches for each participant). We then used these values to compute an average difference score over the participants in each age group (e.g., age 3, 4, and 5) within each random permutation. This averaging process produces 5,000 baseline-derived difference scores for each age group.

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

In each of the plots below, the black dot represents the real data value for the effect being shown (the difference score). The effect sizes from the 5,000 randomly permuted data sets are shown as a distribution. The percentage displayed is the percentage of random permutation difference scores exceeded by the original data differences score (taking the absolute value of all data points for a two-tailed test). Comparisons marked with 95% or higher are significant at the $p<0.05$ level.

Experiment 1: Age and linguistic condition

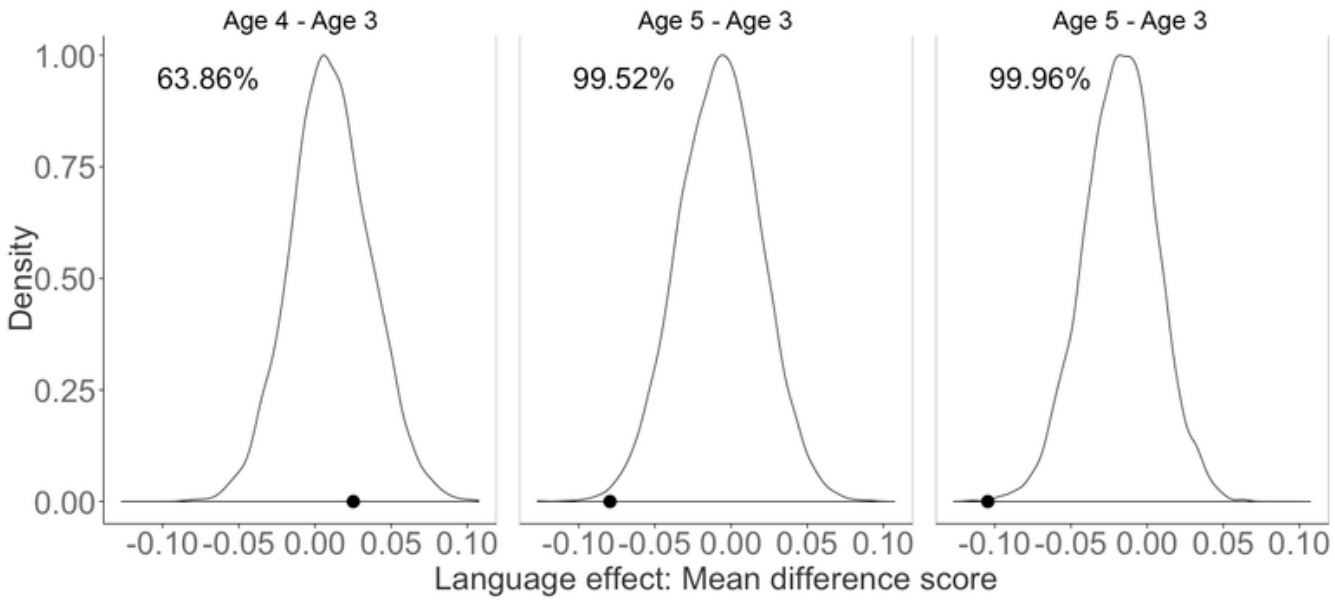


Figure B.1: Pairwise comparisons of the language condition effect across ages in Experiment 1.

Experiment 2: Age and the *prosody only* condition

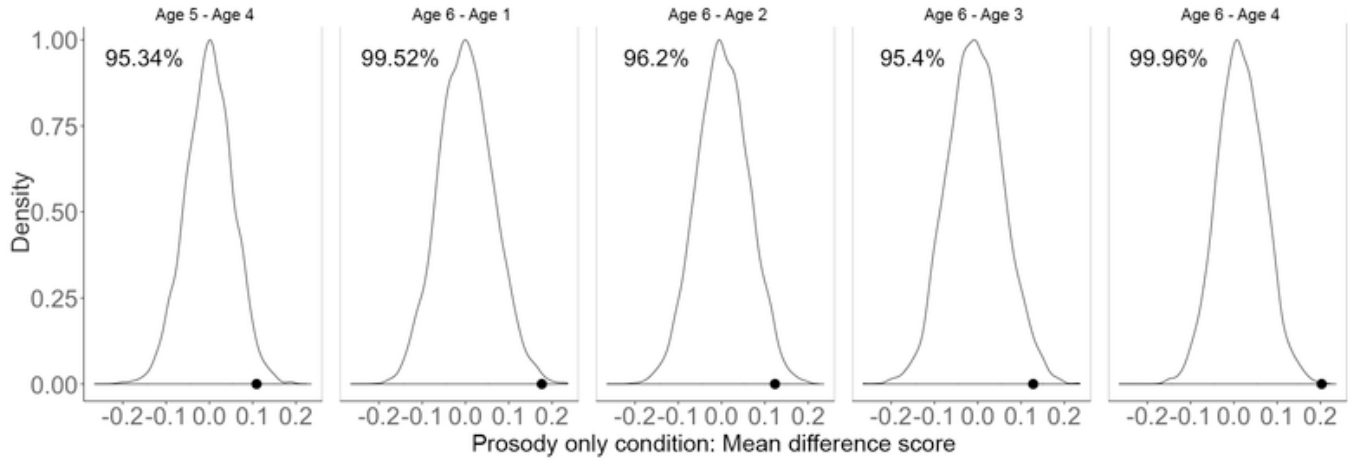


Figure B.2: Significant pairwise comparisons of the *prosody only-no speech* linguisitic condition effect, across ages in Experiment 2. Non-significant comparisons are not shown.

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

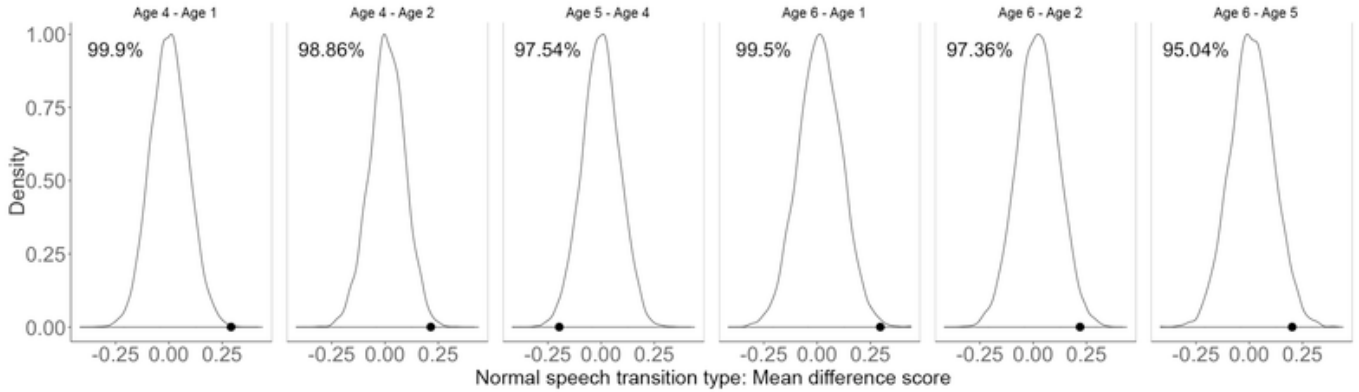


Figure B.3: Significant pairwise comparisons of the *normal speech-no speech* language condition effect for transition type, across ages, in Experiment 2. Non-significant comparisons are not shown.

Material C. Boredom-driven anticipatory looking

One alternative hypothesis for children's anticipatory gazes is that they look at the current speaker at the start of each turn, but then grow bored and start looking away at a constant rate. Even though this alternative hypothesis does not predict the primary effects in our data (e.g., the difference between questions and non-questions), we cannot rule out the possibility that a portion of participants' saccades come from boredom.

The data plotted here show a hypothetical group of boredom-driven participants (gray dots) and participants from the actual data in Experiment 2 (black dots). The hypothetical boredom-driven participants look away from the current speaker at a linear rate, beginning one second after the start of a turn.

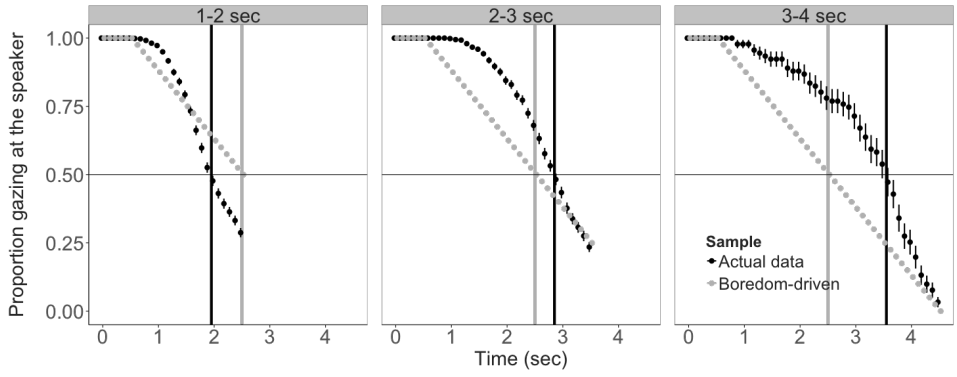


Figure C.1: Proportion of participants (hypothetical boredom-driven=gray; actual Experiment 2=black) looking at the current speaker, split by turn duration. Vertical bars indicate standard error in the experimental data.

If children's switches away from the current speaker were purely driven by boredom, they would switch away equally quickly on long and short turns. Therefore, their crossover point—the point in time at which 50% of the children have switched away from the current speaker—would be the same for all turns, no matter the length of the turn. This pattern is demonstrated in the hypothetical boredom-driven crossover points, which always occur 2.5 seconds after the start of speech (gray vertical lines; Figure C.1).

In children's *actual* looking data we see that crossover points increase with turn duration: 2.0, 2.9, and 3.6 seconds after the start of speech for turns

with durations of 1–2, 2–3, and 3–4 seconds, respectively (black vertical lines; Figure C.1). This pattern suggests that, though children do look away as the turn is unfolding, their looks away are not simply driven by boredom.

Are the looks away in Figure C.1 still too early to count as "turn-transition" anticipation? It is true that children start looking away after one second has passed, but then only gradually. Some of these early looks away may be boredom-driven, but it is equally plausible that some of them are turn-driven. Early predictive behavior is common in turn-taking studies with adults, in both constrained turn-taking tasks (De Ruiter et al., 2006; Gísladóttir et al., 2015; Bögels & Torreira, 2015) and in spontaneous conversation (Holler & Kendrick, 2015; Bögels et al., 2015). Although this same pattern has yet to be established for children's turn predictions, the looking behavior here is at least consistent with adult response patterns in previous work. Additionally, because our analysis windows in the main study only overlapped with the pre-gap utterance by 300 msec (Main paper; Figure 2), our primary results are unlikely to capture any of these very early or early boredom-driven gaze switches, which makes them unproblematic either way in the current analysis.

We therefore conclude that the boredom-driven effects in our data are unlikely to change our primary results, though we acknowledge that characterizing different gaze switching strategies in this kind of data is an important avenue for future work.

Material D. Puppet pair and linguistic condition

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

Methods

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

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

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

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

Results and discussion

We modeled children's anticipatory switches (yes or no at each transition) with mixed effects logistic regression, including puppet pair (robots/mer-people/party-goers/other-3) as a fixed effect and participant and turn transition as random effects. We grouped the red, blue, and yellow puppets together because they collectively represented the puppets used in the *normal* speech condition—this follow-up experiment is meant to test whether the condition-based differences from Experiment 2 arose from the puppets used in each condition.

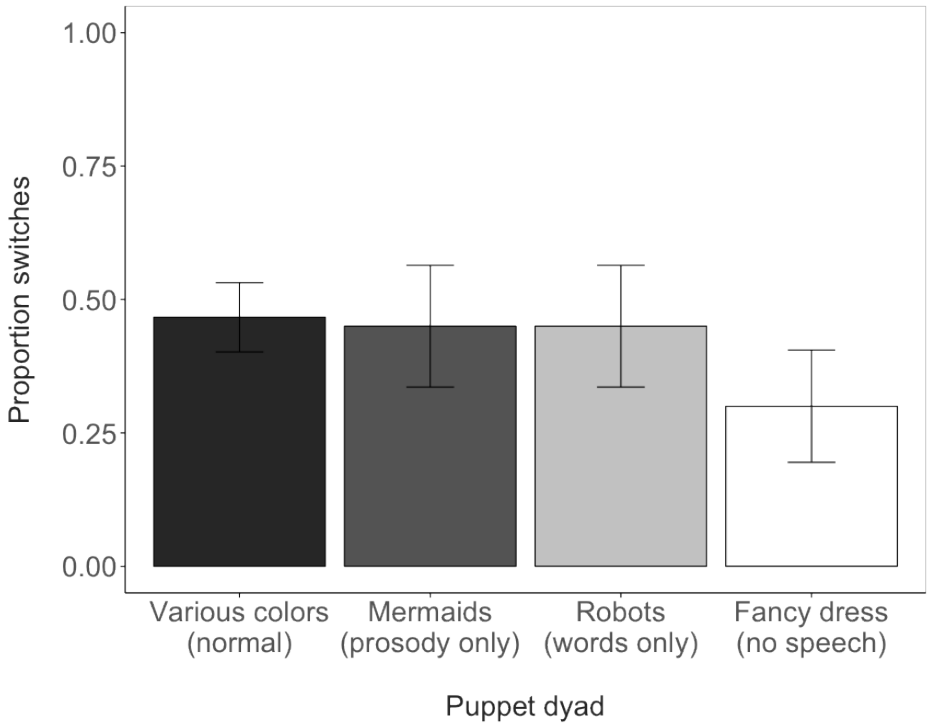


Figure D.1: Proportion gaze switches across puppet pairs when linguistic condition and conversation are held constant.

	Estimate	Std. Error	z value	Pr($>\|z\|$)
Reference level: normal-condition puppets				
(Intercept)	-0.148	0.328	-0.451	0.652
Puppets=*mermaid*	-0.076	0.655	-0.116	0.908
Puppets=*robot*	-0.071	0.653	-0.109	0.913
Puppets=*party*	-0.782	0.687	-1.138	0.255
Reference level: mer-puppets				
(Intercept)	-0.224	0.568	-0.394	0.694
Puppets=*robot*	0.0048	0.801	0.006	0.995
Puppets=*party*	-0.706	0.827	-0.854	0.393
Reference level: robot puppets				
(Intercept)	-0.219	0.566	-0.387	0.699
Puppets=*party*	-0.711	0.827	-0.860	0.390
Reference level: party-goer puppets				
(Intercept)	-0.93	0.607	-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.

In four versions of this model, we systematically varied the reference level of the puppet pair to check for any cross-condition differences. We found no significant effects of puppet pair on switching rate (all $p>0.25$; Table D.2).

We take this finding as evidence that our decision to not fully cross puppet pairs and linguistic conditions in Experiment 2 was unlikely to have affected children's anticipatory gaze rates above and beyond the intended effects of linguistic condition.

References

Allison, P. D. (2004). Convergence problems in logistic regression. In M. Altman, J. Gill, & M. McDonald (Eds.), *Numerical Issues in Statistical Computing for the Social Scientist* (pp. 247–262). Wiley-Interscience: New York, NY.

Allison, P. D. (2012). *Logistic Regression Using SAS: Theory and Application*. SAS Institute.

Bögels, S., Magyari, L., & Levinson, S. C. (2015). Neural signatures of response planning occur midway through an incoming question in conversation. *Scientific Reports*, *5*, Article number: 12881.

Bögels, S., & Torreira, F. (2015). Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*, *52*, 46–57.

De Ruiter, J. P., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, *82*, 515–535.

Gísladóttir, R., Chwilla, D., & Levinson, S. C. (2015). Conversation electrified: ERP correlates of speech act recognition in underspecified utterances. *PloS one*, *10*, e0120068.

Holler, J., & Kendrick, K. H. (2015). Unaddressed participants' gaze in multi-person interaction. *Frontiers in Psychology*, *6*.