We greatly thank the editor and our two reviewers for their very helpful comments on the most recently submitted manuscript. We have done our best to address their comments here and in the manuscript.

1. Are the permutation analyses anti-conservative?

R2 found our new description of the permutation analysis satisfactory, but R1 still expressed some doubts that we believe may have arisen from a miscommunication on our part. We have therefore focused the description of the permutation analyses more on the purpose of the analyses (pp. 32–33). In particular we wanted to address R1's concern that the permutation analyses are anti-conservative. R1 expresses worry that we assume that "baseline saccades are equiprobable across the turn." We have clarified the text to show that this is luckily not the case. Crucially, our permutations assume that baseline saccades are equiprobable across the entire stimulus, and not just across turns. They test the likelihood that we could get significant effects (of any predictor) if the saccades turned out to be randomly distributed—a true random baseline. R1 may imagine that we are comparing (a) anticipation rates at transition windows against (b) rates during turns, thereby maximizing our chances of finding a difference from the baseline. This is not what we are doing and is specifically what we criticize about the Keitel et al. (2013, 2015) studies. Instead, we compare anticipation rates at transition windows to transition rates over the whole stimulus. In other words, our baseline is agnostic to the placement of the true transition windows.

One possible response R1 might have to this is that, presumably, most saccades are at transition windows and so, by comparing them to saccades over the whole stimulus, we are merely comparing our time windows of interest to at a watered down saccadic rate. We have two responses to this objection: First, the primary aim of the permutation analysis is to test the assumption inherent in that idea (i.e., that anticipation at transition windows is greater than elsewhere), but has the substantial added benefit of testing whether the assumption holds for each individual predictor effect in the mixed effect models. Second, the non-"transition window" parts of the stimulus contain many behaviors that likely elicit anticipatory gaze behavior. For example, the transition windows in our real-turn analyses don't include: very early and very delayed anticipatory looks, looks to a responder in error (before the participant realizes the same speaker has actually continued speaking), and turn transitions that weren't included in our final analyses (e.g., because they had overlapped speech or very long inter-turn gaps). All three of these cases are common in our data and elicit similar looking behavior to the target "transition window" anticipatory gazes. But, because they are not part of our final "real transition window" dataset, they only count against us in the analyses—they *only* appear in the baseline estimates, and never in the "real" transition estimates.

We therefore conclude that our permutation analyses are (a) fit to the question we pose, (b) present a more useful and conservative baseline than used in prior literature, and (c) are conservative in the sense that there are other turn-relevant behaviors hidden in the non-"transition window" parts of the stimulus. By editing the motivation behind the analyses, we hope we have made this argument clearer for readers. We hope too that this satisfies R1's concerns. He or she may still be somewhat unimpressed with the permutation analyses becauuse they do not capture anything more than "random-looker" vs. "turn-driven looker" (in his/her words, that they are "superfluous"), but this is exactly what they are intended to do. We do think that R1's point about boredom-driven looking is interesting, however, which is why we test this more subtle hypothesis about anticipatory looking rates (i.e., "boredom-driven looker" vs. "turn-driven looker") in a separate set of analyses (Boredom Hypothesis; Appendix C, also discussed below).

2. Are some of the saccades still driven by boredom?

R1 notes that, although we show significant differences between our real lookers and hypothetical "boredom-driven" lookers in Appendix C, our additional analyses do not rule out the fact that boredom might play a role. R1 points to the Figure C.1 to highlight that children still sometimes look away from the speaker long before the turn ends. We completely agree that our analyses can't rule out the fact that *some* proportion of the saccades might be due to boredom (or to chance). Surely R1 is right that some of them are! That said, we do not think that this is a major issue for our conclusions. First, the primary effects in our data (which revolve around questions and access to linguistic cues) are not predicted by a boredom account—a bored looker should be constant in their boredom for questions and for non-questions. Second, our data show dramatic difference in the timing of look-away rates across turns of different length, which is also not predicted by a boredom-based account: boredom-driven look-aways would occur at a constant rate, independent of the ongoing turn's ultimate length. Finally, early (within-turn) predictive behavior is a normal finding for turn-taking studies. The bulk of previous work with adults, in both constrained turn-taking tasks (e.g., de Ruiter et al., 2006, Gísladóttir et al., 2015; Bögels et al., 2015) and spontaneous conversation (e.g., Holler & Kendrick, 2015; Torreira et al., 2015) have found evidence for early speech act detection, early response planning, and early anticipatory looking, all occurring well before the end of the ongoing turn. We therefore believe that whatever boredom-driven effects exist in our data are likely small and do not interfere with the interpretation of our primary variables of interest. We have added discussion on this point to the Appendix section dedicated to the boredom hypothesis (pp. 108–109). We also added a footnote earlier in the manuscript, pointing readers to this extra analysis in case they are also interested in this follow-up (p. 37).

3. Are the permutation-based t-tests of age effects trustworthy?

We appreciate that our use of the random baseline data to compute likelihood is an unusual approach. We chose it to fit the particular analysis needs of these studies. That said, R1's comments encouraged us to include more traditional tests to supplement these analyses (and perhaps thereby make the findings more comprehensible overall). We have therefore added standard (non-permutation based) comparisons between the Normal speech and No speech ("party") conditions for each age and have integrated them into the results for Expt 2 (pp. 57–58).

4. Shouldn't the effect of gap duration be part of the developmental story?

R1 suggests we should treat gap duration as more than just a control variable; that its effect is both strong and theoretically interesting. We think this is a really great point. Following R1's suggestion, we have added discussion about the role of inter-turn gaps for retroactive turn boundary assignment throughout the manuscript, from the Introduction to the General Discussion. We have worked to make it clear why gaps themselves can be a good cue to turn transition. We have also taken up R1's suggestion for reframing in the General Discussion; we now discuss the developmental trajectory with respect to both the condition-based linguistic effects and inter-turn silence, which more naturally integrates the effect of gap duration into our story. By doing this, we move "away from the idea that linguistic prediction is the only explanatory factor" as R1 proposes. R1 also noticed that we didn't include an age-by-duration or condition-by-duration interaction in our models. That was because we were unable to add full interactions between gap duration and the other fixed effects in all models; in many models, higher-order interactions with gap duration create non-convergence issues due to data sparsity with some combinations of three or more predictor levels. We carefully considered how we could satisfy R1's comment while also retaining consistency and validity in our models.

Our analyses now test the effect of gap duration but also remain consistent across all of our models. We achieved this by adding a fixed effect of gap duration and two-way interactions between gap duration

and all other simple fixed effects (i.e., gap duration and age, linguistic condition, and transition type) to every model. This way all of our models test both the overall effect of gap duration (as before) and any interaction between gap duration and age, linguistic condition, or transition type. This tests the hypothesis that the gap duration effect might change with age that R1 put forward (which, incidentally, we do not see in our data) as well as other potential interactions between gap duration and other predictors.

5. Is there strong evidence that children are using linguistic information?

R1 mentions that the linguistic effects in our study are relatively small, all tucked within interactions. R1's comment is correct, but we would also like to point out that we expected from the start that linguistic effects would strongly interact with age (the reason we used a broad age range) and that anticipatory gaze would interact with speech act (the reason we included question coding). We did not, however, anticipate that the effect of speech act would so strongly dominate our results. The strong effect of speech act effectively funnels all the linguistic effects (and their age interactions) into the "question" subset of the data. Because the speech act effect is both new and theoretically motivated, and because there are still effects of age and linguistic condition associated with it, we believe that these findings are informative: they tell us about the conditions under which children and adults engage in spontaneous linguistic prediction about turn taking in conversation.

To better make this argument in the manuscript, we have added discussion on the origin of the question effect and why it is important (also addressing a request from R2 below). R1's comment may have partly arisen from the way we discussed the linguistic effects. We reviewed the text and found that, indeed, we occasionally overgeneralized linguistic effects to "children" (as a whole) when they really should be limited to "older children" or "adults". We have gone back through the text and made sure that, when discussing linguistic effects, it is clear which subset of participants we are referring to.

R2 points out some conflicting statements about the use of lexical cues in Experiment 2. Thanks! These are now addressed in the text. R2 also urges us to consider our findings with respect to other cues that might drive anticipation (e.g., non-verbal cues). We have added a short point to the discussion that makes clear that the linguistic effects we study here are only a few of the possible sources for anticipatory information, and that the systematic study of other cues, especially non-verbal ones is still needed (p. 64).

6. Is the developmental story too focused on the condition-based linguistic cues?

R1 proposes that it is an oversimplification for us to argue that "turn prediction develops in early childhood", and that a better story would highlight two components: (a) reactive cues (e.g., inter-turn gap duration) that children may use early on while they (b) gradually learn the association between certain linguistic structures/speech acts and speaker transitions (for which they must use linguistic cues). We thank R1 for this suggestion! As mentioned above, we have integrated discussions of gap duration as a cue to turn taking throughout the text, outlining this more nuanced developmental trajectory in multiple places.

7. What is the source of the question effect?

R2 asks us to speculate about the causal explanation for the question effect, particularly regarding whether it is really about speech acts or about other effects that correlate with questionhood (e.g., predictability). He or she also raises the possibility that the question effect might have developmental roots, e.g., from the high frequency of questions in children's input. We thank R2 for the great comment and have added some discussion concerning these points (p. 67).

8. Should we be concerned about these low rates of spontaneous anticipatory looks?

R2 suggests that the low rates of anticipatory gaze switching in our data may cast doubt on the idea that adults and children routinely anticipate turn ends, as is assumed in much theoretical and experimental work about the psycholinguistics of turn taking. If so, R2 would like us to comment on how, then, speakers still successfully participate in conversation. We have added a brief note on our overall anticipation rates to the text of the General Discussion (p. 68). Because this is a somewhat niche question that most readers may not be interested in, we only respond more elaborately here. But we thank R2 for encouraging us to think more about how our findings fit in with others on turn prediction:

In our study, "spontaneous prediction" is a simplified indicator that participants successfully anticipated a transition from the prior speaker to the upcoming speaker with their eye movements. The absence of an anticipatory gaze, does not signal that no anticipation was made. Behavioral evidence from experiments with adult listeners (e.g., Bögels et al., 2015; Sjerps & Meyer, 2015) also suggest that participants can flexibly calibrate the onset of their response planning, such that they may pay more or less attention to the turn-final content of incoming turns. There is no reason to think that children would, unlike adults, always use the same predictive strategies. Therefore, in some sense, we do see our rates of anticipation as challenging for the idea that participants always or even usually make precise prediction about upcoming turn ends, but our findings are not the first to support that idea. What does this mean for first-person prediction?

Other studies using similar methods in third- and first-person measures find very similar anticipation rates, so our findings are at least not anomalous (e.g., Keitel et al., 2013; 2015; Holler & Kendrick, 2015; Hirvenkari, 2013). Our data also primarily tap into an effect that should generalize to a first-party situation: the question effect. The question effect might be amplified when participants themselves need to respond. We briefly address these points in the text (p. 68).

In sum, we think there is still good reason to believe that participants routinely anticipate upcoming turn structure (e.g., the ends of ongoing turns and the likelihood of upcoming turn transition), but our study does run counter to the idea that they do this spontaneously at all possible points of turn transition. Our findings are in-line with an account that emphasizes flexibility in how and when participants make predictions about turn taking.

9. Do the multilingual children behave differently?

R2 makes the intriguing suggestion that the subgroup of participants who have exposure to a second or even third language might behave differently because of their more varied experiences with language and turn taking (particularly with respect to turn-relevant cues such as prosody). We agree that this is a fascinating question, but we will not analyze this subgroup separately for a few reasons: first, as the reviewer guesses, we would not have enough statistical power. Second, and perhaps more importantly, we did not conduct a full language history of each participant and so we do not know whether the second language comes from parent(s), grandparents, a nanny, friends, school, etc., and therefore can't confidently group children into a uniform "multilingual" group. We do think this is an important point, though, so we've added a footnote on this point (p. 17).

10. Does speech style affect children's performance?

R2 wonders whether our results imply that children's performance in our study is better than in previous, similar work (i.e., Keitel et al., 2013, 2015) because we use child-friendly speech in our stimuli and, if so, why that might be. He or she suggests that the mechanism could be linguistic, attentional, or both. We agree that it is likely both linguistic and attentional, and have added a note about this in the manuscript (p. 70).

11. Are the non-English "question cues" comparable to English question cues?

R2 correctly points out that the status of the prosodic cues in the non-English recordings from Experiment 1 is not clear because, though they might be recognizable by English speakers as "similar to" English question cues, they are still phonologically and phonetically not identical to actual English question cues. We agree, but we also think that this doesn't bear too heavily on our findings. However, to do justice to R2's comment and to make sure that this fact doesn't pass readers by, we now acknowledge the difference the non-English prosodic cues more emphatically (p. 20).

12. Can you better emphasize the strengths of this method?

R2 would like us to more explicitly mention the benefits of our method for measuring anticipation. We thank R2 for his or her encouragement and have added a this to the General Discussion (pp. 73–74).

13. Is it important that this is a third-party predictive measure?

The editor would like us to comment on how participant role might affect our results. We link this request to R1's comments about the use of gap duration for predictive and reactive turn tracking, and to R2's comments about the generalization of our anticipation rates to conversation more generally. As discussed above, we have taken some space in the General Discussion to speculate about the possible effects of measuring predictions about turn taking from a first- or third-person perspective (e.g., pp. 68–69 and pp. 72–74).

14. Miscellaneous issues with the graphs

Both reviewers made small requests regarding the graphs. In line with their requests, we have removed the legend for condition (color) in Figures 3 & 6, matched the labels for "question" and "non-question" in all cases (i.e., not "Q" and "S" as they sometimes still appeared). We also made Figure 2 clearer with examples of the earliest and latest possible anticipatory gaze switches, plus a prototypical anticipatory gaze switch.

15. Other minor issues

We addressed a number of minor issues, including removing internal numbering, using APA-formatted references, adding a brief note about the non-centered age variable for each experiment, and some smaller comments on confusing wording and typos.

In making these changes, we feel that we have substantially improved the clarity of our developmental story and our theoretical and methodological contributions. We give big thanks to the editor and reviewers again for their insightful and helpful comments and hope that they also find the manuscript to be much improved!