We thank all three reviewers for their thorough and helpful comments on the previously submitted manuscript. We have tried to address all of the reviewers' comments, within the manuscript and/or here in the response letter.

Reviewer 1

1. 333ms may be too long for older children's saccadic planning time and may therefore inflate the anticipatory looking rates for older children.

We were concerned about this, too, but we unfortunately did not collect saccadic planning norms for children ages 1–6. We did not feel comfortable (semi-)arbitrarily assigning saccadic planning times to different ages, so in the newly analyzed data we take the conservative approach of using adult-like planning times (200ms) for children at all ages in both experiments.

2. The description of the random baseline permutations was not clear.

We have tried our best to clarify this description (Section 2.2.2) and have added specific pieces of information where requested.

3. Are anticipatory gaze shifts driven by boredom (and not by anticipation)? Do children generally just look away as the speaker goes on?

This alternative hypothesis makes a valuable test of what we measured in the experiments. In the Appendix you can now find a graph comparing theoretically boredom-driven lookers and our real lookers. We modeled the proportion of participants looking at the current speaker for our real data and for (fake) boredom-driven lookers (who look away from the current speaker at a constant rate, starting 1 second after the onset of speech) across turns of different length. Even in short turns, where there is very little temporal room for children to behave differently from the boredom-driven lookers, there is a clear difference between the boredom-driven and the real data. (Figure D.2). We take this to mean that children are unlikely to be looking away simply due to boredom.

4. The reported results form the models models are hard to follow.

We have added a little more information about how each variable was coded and have put summary tables of every model's output into the text.

5. "No speech" should be the reference level against which the linguistic conditions are compared in Experiment 2.

This is a very useful comment, thanks! We agree and have changed our analyses to accommodate it. We had started with the "normal speech" condition as the reference level because that is what was done in related work on adult turn-end prediction. In the end, however, we too preferred the additive model, and have changed the text accordingly.

6. Looking at the data, it seems possible that the children are not using any linguistic information.

Thanks for asking for clarification here. We have tried to integrate aspects of our answer here into the manuscript. The lack of simple language condition effects in the children's data in Experiment 2 does not imply that they were not using linguistic information at all in their predictions (i.e., solely relying on visual information in the stimuli and the duration of inter-turn gap). An unexpected but quite important take away of our findings is that children and adults are far more likely to make an anticipatory gaze switch after hearing a question. Because speech

act seems to be so effective in driving participants' anticipations, the extent to which they use linguistic cues is probably limited to their use of those cues to identify questions in the ongoing turn (which in turn allows them to expect an upcoming speaker switch with high certainty). The children's data showed a 2-way effect of age and transition type (question vs. non-question), but only in the "normal" speech condition, suggesting that it was the only condition in which they could extract the linguistic cues needed to accurately identify questions, perhaps because prosody and lexical cues to questionhood often work together in forming questions during everyday speech, but perhaps because the manipulated speech was simply too unfamiliar for them to apply their normal interactive prediction processing.

Either way, the existence of the question effect in the normal condition indicates that children do, in fact, use linguistic information (to identify questions) and the interaction of that effect with age suggests that they get better at doing so as they get older. These effects, for example, cannot be explained by duration alone (and its effects are already accounted for in the statistical analyses), and there is no visual cue advantage difference between the normal condition and the others. We are then left to conclude that they use linguistic information to support the question effect. A similar story can be told for the data in Experiment 1: non-verbal cues to questions vs. non-questions are available in both the English and non-English stimuli, but participants (adults and children) made more anticipatory switches when they had access to the lexical information. Again, this cannot be explained by gap duration or unequal visual advantages across the stimuli.

The fact that we have found most of our linguistic processing effects within an effect of speech act paints a very different picture from prior work (e.g., de Ruiter et al., 2006), which finds much clearer evidence of linguistic processing across multiple phonetically manipulated conditions, but without question-non-question comparisons. No turn prediction measure yet is capable of capturing predictions that occur in spontaneous first-person conversation, but we think that our results indicate that there are multiple types of prediction relevant to anticipating upcoming turn structure during interaction, and that any account of how listeners use linguistic cues while participating in conversation should be able to account for shifts in how listeners attend to cues to make more or less specific predictions (i.e., "what words are upcoming" vs. "what types of action are upcoming").

In sum, while the linguistic effects in our data are somewhat limited, they are present, consistent, developmentally changing, and important to our developing theories about turn prediction. We believe that this sets up future work quite well to dig into the specific linguistic cues (lexical and prosodic, though probably mostly lexical) that are being used in spontaneous question identification.

7. It is too hard to link the conclusions in the discussion back to the statistical results and the effects themselves are not talked about in a clear enough way.

We have tried our best to make this clearer, first by clarifying the statistical analyses and outputs, and second by trying to explicitly link each claim to an individual result (through the use of figure and table references). We hope that this will help readers better connect the results to our interpretations.

8. There are too many references to unpublished work.

The single upside to taking so long in implementing these comments is that all of the unpublished work referred to previously is now published or in press!

(Reviewer 1's minor points/corrections have all been addressed directly in the text.)

Reviewer 2

1. It is not clear which claims are backed up by which analyses. Analyses should be carried out for each group separately.

Thanks! We have tried to improve the linkage between our claims and analyses and now have presented the statistical results in tables, as requested. Please see responses 4 and 7 under Reviewer 1 for a little more detail. The main focus of our analyses was the effect that age (as a continuous variable) has on anticipatory gaze. However there are two cases in which our main analyses are not satisfying without further, age-group analyses: (a) age sometimes showed itself to interact with other factors (e.g., age and language condition), and (b) without further tests we can not say whether particular age groups themselves statistically differ from chance, across or within conditions. Following your suggestion, we have therefore done two things. First, we added follow-up two-tailed t-tests, making pairwise comparisons between age groups for significant interactions so that we could find out how the interacting factor (e.g., language condition) significantly changes across age groups (Sections B.1–3). Second, we have added individual models of the youngest age groups (ages 1, 2, and 3), comparing their real looking behavior to our randomly permuted data across conditions (at the ends of Sections 2.2.2 and 3.3.2). These follow-up analyses, which are summarized in the text, are primarily described in the Appendix (Section B).

2. The results and discussion for each study would be better if they were separated into different sections.

Our original intention in combining results and discussion was to make the findings easier to understand. Since we were not able to achieve this in the prior submission, we have followed Reviewer 2's advice and now use a more traditional results-only then discussion-only structure. We think that this has improved the clarity of the paper and we give thanks for the suggestion.

3. The methods sections need more elaboration in a few places, and the baseline analysis section could be much clearer.

We thank the reviewer for his or her careful attention to our methods section (and its missing details.) We have added in the experimental methods information requested and have re-checked each section for missing information someone might need to replicate our methods. We've also worked to clarify the description of the baseline analysis, including a new figure to accompany it (Section 2.2.2; Figure 4). We hope that it is now easier to understand.

4. The adult data deserve further discussion because they raise the question of whether this measure taps into linguistic processing at all.

We actually think the adults' results are well in-line with prior work showing that (a) adults' predictions primarily rely on lexical cues (E1 & E2) and (b) adults are better anticipators than young children (E1 & E2). The complicating factor is that we report a new and important interacting factor with their linguistic processing: speech act. Like the children in our data set, adults made more anticipatory switches following questions, which means that most of our evidence about their use of linguistic cues comes from question transitions. This enriches the picture of linguistic processing for turn prediction, as painted by prior work: linguistic processing is brought into play more often in some contexts than others. From these points we conclude that our experiments do indeed tap into participants' use of linguistic cues to predict upcoming turn structure. That isn't to say that the method wouldn't be dramatically improved with more

natural linguistic controls and tighter controls on gap duration—it would! But we think our current results still make a valuable contribution to the current understanding of turn anticipation in children *and* adults, and lays critical groundwork for discovering what specific types of cues participants spontaneously monitor in everyday interaction.

5. There is a lot of unpublished work referred to in the manuscript, with errors in the reference list.

These papers are now all published or in press. The errors in the reference list were partly due to half-filled BibTex entries used to produce the previous manuscript. We have checked them again and hope that there are very few (or no) errors remaining.

Reviewer 3

1. The manuscript could be shortened substantially without loss of content.

We agree that the paper was overly long. We have significantly shortened the introduction and tried to remove redundancies where possible in the rest of the text. We have also moved some of the secondary analyses to the Appendix so that the main text can be read without getting into the nitty gritty of every single analysis. We hope that this has helped make the main text more straightforward and approachable length-wise.

2. There is a confound between linguistic condition and puppet pair in Experiment 2.

We thank Reviewer 3 for this point. We repeat here some of the information now available in the Appendix for easy summary: Our design 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, or pancakes), robots were only associated with "words only" speech, merpeople were only associated with "prosody only" speech, and the puppets with fancy clothes were only associated with the "no speech" condition. We did this to increase the pragmatic felicity of the experiments for the older children (i.e., robots make robot sounds, merpeople's voices are muffled under water, the fancy-clothed puppets are in a room with main other voices). It is therefore fair to point out a possible confound between linguistic condition and puppet pair. Thankfully, we also ran a short follow-up study at the museum with 3–5-year-olds in which each child only saw one video—the normal speech conversation about birthday parties—with a randomly assigned puppet pair performing the conversation. Five children watched each puppet pair, for a total of 30 children across the six pairs (shown Figure 3). This experiment holds all things constant except for the appearance of the puppets. We then used a mixed effects logistic regression of children's anticipatory switches (yes or no at each transition), with puppet pair (robots/merpeople/fancy dress/normal-speech-puppets) as a fixed effect and participant and turn transition as random effects. In four versions of this model we systematically varied the reference level to check for differences between every puppet pair, finding no significant affects of puppet type on switching rate. We take this as evidence that, although we did not fully cross puppet pairs and linguistic conditions in Experiment 2, it was unlikely to have had strong effects on children's looking rates above and beyond the intended effects of linguistic condition. We have included details and graphs for this analysis in the Appendix.

3. Computer animation would have been a better choice than puppets.

We completely agree with the reviewer on this point. We did not have the ability to create custom animated conversation videos at the time. But, since then, the first author has developed a method for doing exactly that. Unfortunately, for this past work, we are still stuck with the

puppet stimuli.

4. The model design needs to be further justified and the way the effects are talked about needs clarification.

We corrected small mistakes in the model description (thanks for spotting them!), for example, making it clear that we used a logistic regression (for our binary response variable of switch-no switch) and have added short bits of information justifying our use of separate models for adults and children and our use of random slopes. We have also cleaned up the way we talk about the significant effects in the model output because they were sometimes misleading. There was no concern about dummy coding since this is carried out automatically across factor predictors by the glmer package in R (as is now clearer with the presentation of results in Tables 2 and 5).

5. An ideal-observer analysis of the usefulness of the linguistic cues would give a better idea about what can affect anticipatory gaze.

We really liked this suggestion but felt that it was too far outside the scope of the current paper, especially because our stimulus set is probably too small to build a generalizable model of cue usefulness. We will, however, seriously consider this type of analysis for future work, with adult and child turn-taking, in experiments and in spontaneous speech corpora.

6. The random baseline would be better if it were converted to a full permutation analysis so that the baseline rate of anticipatory gazing is not treated as an additive factor, which it probably is not.

We thank Reviewer 3 very much for his/her suggestion and clear explanation of how to conduct the analysis. To be sure about the results we decided to run 5,000 random permutations on the data from each experiment, resulting in 20,000 glmer runs. We apologize for the delay of this resubmission, which is primarily due to the implementation of this random permutation analysis. To analyze our data on this scale, the first author had to learn some new R libraries, optimization techniques, and how to use the local computer cluster, dealing with many technical problems along the way! That being said, we think that the new permutation analysis is indeed simpler and more theoretically intuitive than our prior method (described in 2.2.2 and 3.3.2, visualized in Section A). Please note that there were two problems with conducting the analyses as originally instructed in the reviewer's comments: (a) we used t values instead of β estimates because the standard error estimates for the randomly permuted data were nearly categorically higher, dramatically so in some cases, and (b) many of the models (especially in Experiment 2) resulted in convergence warnings, which we have dealt with by only reporting results for converging models.

7. The theoretical contribution of this paper is not clear.

We have tried to improve the clarity of our findings' theoretical importance in the revised manuscript. I summarize here the points that we cover. One of the most important theoretical contributions of our work is in establishing and replicating speech act effects in children and adults' spontaneous predictions—while participants may always use linguistic information to predict upcoming speaker changes, our results to not support the idea that they always use linguistic information to predict upcoming turn ends, as is assumed in metalinguistic measures of turn-end prediction (e.g., pressing a button while listening to speech). Our results suggest that participants' spontaneous predictions are the result of question-monitoring, presumably achieved by recruiting linguistic cues to questionhood from the unfolding signal. The trends in our data suggest that participants are primarily recruiting lexical cues to do this, though establishing this pattern will require follow up work with more focused stimuli.

A second, but related theoretical takeaway is that lexical information alone is not equivalent to full linguistic information for children, as it has been shown to be (e.g., de Ruiter et al., 2006) and replicated here (Tables 2 and 5) with adults. If this effect arose in our data because children do not respond well to phonetically manipulated speech in conversational contexts, this point can be overturned by work that controls linguistic information through other means. If not, it suggests that there is something specially informative about the combined prosodic-lexical marking on question turns, for which there is also prior support (Torreira et al., 2015) though our finding would be the first to show such effects developmentally.

A third theoretical contribution is that young children (e.g., at ages one and two) do make anticipatory gaze prediction more often than would be expected by chance alone, but not by much. Since older children and adults' gaze shifts are primarily driven by question turns, this may suggest that, although children *can* make predictions about upcoming turn structure, it doesn't count for much until they have acquired the linguistic skill to pick out question turns. This finding, in turn, has potentially important implications for how participant role (first- instead of third-person) and cultural differences (high vs. low parent-infant interaction styles) might affect children's early predictive skill. It also bridges prior work demonstrating that children begin taking non-linguistic turns in infancy (e.g., Hilbrink et al., 2015), but still have trouble integrating linguistic aspects of responding at age 3–4 (Casillas et al., in press; Garvey & Berninger, 1984).

The fourth contribution is in our interpretation of discrepancies between our methods and results compared to prior work (using more explicit, metalinguistic methods and showing more obvious online predictions using linguistic information). The analyses of these discrepancies led us to put forth a new idea: participants in first-person spontaneous conversation may use multiple strategies in making predictions about upcoming turn structure, using more passive prediction to detect upcoming speaker transitions (e.g., questions), switching into precise turn-end prediction mode when necessary. A flexible system like this allows listeners to continuously monitor ongoing conversation at a low cost while still managing to plan their responses and come in quickly when needed. As far as we know, we are the first to articulate this idea, integrating findings from multiple methods, age groups, and styles of linguistic control.