Dear Dr. Ouellette,

Thank you for the opportunity to revise and resubmit this article with coauthor Michael C. Frank entitled “Young children’s developing sensitivity to discourse continuity as a cue for Inferring Reference” to Journal of Experimental Child Psychology. The questions, concerns, and suggestions set forth by you and the reviewers were very useful in helping us focus the paper and in repackaging our

results. Below is a detailed list of revisions that we have made in response to the individual points raised.

We look forward to hearing from you.

Sincerely,

Alexandra C. Horowitz & Michael C. Frank

**Reviewer #2**

*Title/Abstract/Highlights: These sections clearly describe what the authors have done. They may overstate the conclusions that can be drawn from this project (see comments in Results & Discussion below). The Title and Abstract (and page 3, line 23) mention word learning, but much of the information in the paper refers more specifically (and accurately) to meaning learning. The writing here, particularly in the Abstract, is somewhat overdone and jargon-y. Clarity and simplification would improve the readability of this section.*

We have corrected the title and reference to describe inferences about meaning rather than necessarily word learning. The abstract was edited to introduce the design more clearly.

*Introduction: The structure and logic of the introduction do not clearly lay the theoretical groundwork for this project. While the introduction appropriately describes the much of the relevant literature, the links between ideas is not always fully fleshed out on the page. Many ideas get raised, but don't get connected to the idea which follows, or the implications of those findings with respect to the current study are not fully described.*

We have edited the introduction to make these connections more apparent.

*For example, at the bottom of page 3, in describing how children would succeed in inferring the meanings, there is limited theoretical context for understanding why you would or wouldn't expect children to be able to make these inferences although it is posited that it would "presuppose a substantial amount of knowledge" (page 4, line 7). Where would this knowledge have come from? The subsequent paragraph goes on to describe adult abilities, but it's not clear how those abilities are acquired. (Some consideration of the statistical learning models might be relevant here?)*

This paragraph has been edited to setup how this argument relates to our example within the broader framework of the introduction’s discussion of adults’ and children’s processing of discourse information.

*While I am fine with the simpler operationalization of discourse that your use in your study, the text seems to move between this simple definition and a more complex one without always making these transitions clear. In addition, given that this paper wants to make claims about the utility of discourse continuity as a clue for meaning learning, the difference between real-world discourse continuity and the simplified version investigated here needs more expansion.*

We have made the use of our operationalization more explicit and consistent throughout the introduction to reduce this confusion.

*On page 5, at the end of the first paragraph, you imply that the Samuelson & Smith (1998) argument is an important one, but then the issue is dropped. How is "the possible role of attention and memory factors" relevant for the present study? How are the Akhtar et al., (1996) findings relevant? These connections need to be more explicitly drawn.*

This argument has been elaborated.

*Overall, the introduction has most of the necessary pieces, but the connections between them and to the argument made in this paper were missing.*

We have tried to make these connections clearer.

*Figure 1 was extremely helpful but seemed somewhat redundant with the information in Table 1.*

We appreciate the reviewer’s comment. While we agree that this information is partially redundant, we believe that including both the figures and the table with the scripts will ultimately be beneficial to help clarify a complicated design to our readers.

*I would have liked to know what all the unusual toy items were, particularly given the disconnect between the images in Figure 1 and the examples provided in the text on page 9, line 34. (And an issue raised in the General Discussion, see below.)*

We now include images of the toys used in Figure 2.

*The inclusion of both study's information in Table 1 (and in the Results section below) was a bit confusing as it necessitated flipping back and forth from the text, although the side-by-side comparison was helpful for Table 1. I will admit that I made some (negative and subsequently incorrect) assumptions about the design of Experiment 2 because I saw this information before I read the text which outlines the rationale of the design; this may have been a quirk of my reading or it may reflect some deeper structural issues.*

Thank you for noting this. Although we would like to avoid misleading predictions about each of the experiments, we believe that putting the scripts side-by-side for direct comparison will be more useful for readers to understand the subtle differences between the design of Experiments 1 and 2,

*Given how the participants were recruited (and my anecdotal observation of families' repeated visits to the local science museum), is there a possibility that any participants were in both studies? Novel toys and words would no longer be so and, given the counter-balancing of toys, words, etc., second-time participants could have confusion from previous exposure.*

Thank you for noting this. We clarify that participants were uniquely identified by their birth dates and verified with video records.

*Results: There are some misstatements about the findings and lack of explicit connection between the predictions and analyses.  
  
Experiment 1: The first sentence includes a misstatement; not all children in the student were more likely to select the 'correct' toy, 2-year-olds failed to demonstrate this finding.*

We have clarified that this was an “overall” finding and elaborate that the youngest children did not reliably succeed at the task.

*Given that direct comparisons between the results of Experiment 1 and 2 are not relevant, the choice to combine the results in a single figure (and tables) was confusing. It invites (and suggests) comparison, but none are explicitly made and there are no statistical comparisons of these findings.*

**PRESENT THESE SEPARATELY?**

*On page 13, line 37, the authors omit reference to the adult findings here. Based on Figure 2 and Table 3, they're clearly consistent with the older children's results, but including them in the text is necessary.*

Adult results are now explicitly included.

*Experiment 2: The predictions for this experiment seem somewhat different than for the first experiment, but the statistical analyses haven't changed. The specific rationale for these analyses and the interpretation of them in this section is weak. While you do point out that there are no significant differences between After First Toy and After Second Toy trials, this appears to be counter to your prediction in the introduction for this section, but this may be a reflect of lack of explicit connection between these analyses and the predictions.*

We have tried to make the rationale clearer. Although predictions are different for the trial types in this Experiment, our model should reveal response pattern differences if present (and in which case we could run more follow-up analyses). However, we found that participants did not respond differently across trial types or from chance in Experiment 2, indicating that temporal proximity did not seem to be playing a substantial role in reference selection.

*Discussion: The conclusions about Experiment 1 are clear and supported by the data, but those for Experiment 2 are more ambiguous, although the text doesn't suggest this interpretation. The data feels overinterpreted. While I agree the findings are suggestive, the text implies a clearer result than the statistics support.*

Our general discussion now more explicitly describes our motivation and predictions for Experiment 2, and how our findings fit in with our interpretation.

*One interpretation given to the weaker findings for the After Second Toy trials is that participants might be constructing a higher-order category. This interpretation is dismissed because the toys are "sufficiently dissimilar," but because full descriptions of all toys haven't been provided, readers must take your word for this. In addition, this is an adult interpretation of what category might mean; we don't know how young children might interpret or create this particular category, so this claims seems unsupported. The creation of an overall category seems much more intuitively connected to a condition which wasn't run, Before First Toy.*

We have added photos of the toys for reference to the reader. Because each item differs by form and function, made explicit through the experimenter’s comments and demonstrations of each toy’s unique affordances, we believe it is unlikely that children would assume them to be in the same category. However, if children did associate toy pairs to a single higher-order category, we would expect to see chance level performance in Experiment 1, but instead we found that systematically selected the toy by naming location with age.

*The logic of the argument on page 17, lines 54 through page 18, line 14 is not clear. More of the rationale needs to be on the page.*

This argument has been elaborated.

*In the final paragraph, the distinction between the simplified definition of discourse and a broader theoretical one is again muddied, suggesting that the results may mean more than the data support.*

The paragraph has been modified to explicitly comment on the contributions of this work before introducing a broader theoretical framework.

*References: Horowitz & Frank, 2013 is included in the reference list, but is not included in the text. There are several incorrect inclusions of issue numbers throughout the reference list.*

Thank you for noting this. The Horowitz & Frank (2013) reference comes from the title page where we note that, “An earlier version of this work was presented to the Cognitive Science Society in Horowitz and Frank (2013).” The issue numbers have been corrected.

*Grammar/Language/APA style:  
p. 5, line 49: 6-18-month-old children (there is a hyphen missing after 18)  
p. 6, line 35: the referent of they is ambiguous. (I assumed it was Frank et al., 2013, but that was half a page and more than a paragraph ago.)  
p. 9, lines 14-18 & p. 15, lines 41-46: you appear to be using en dashes where you should be using hyphens (2-year-olds, etc.)  
p. 9, line 16: 3-year-olds years--the word years should be struck  
p. 9, line 16 (and various other locations): = signs should have a space on either side (e.g., n = 16, not n=16)*

These have all been corrected.

**Reviewer #3**

*3.      The title is appropriate apart from the use of the term 'word learning'. The authors acknowledge in a footnote 2 on page 6 that children do not necessarily retain words following in-the-moment referent selection. Referent selection and word retention are two distinct processes and the line between the two are sometimes blurred (see Spiegel & Halberda, 2011). This is particularly the case when only one of the objects is named and children are not required to disambiguate objects on the basis of the names, but only which one was linked with a novel name (Dollaghan, 1985; Schafer & Plunkett, 1998; Wilkinson, Ross, & Diamond, 2003). This is an article about children's inferences about a speaker's intended referent of a novel word rather than word learning per se. Therefore the title of the article might need to be changed by dropping "word learning" to something that more accurately reflects the tasks, such as 'naming inferences' or 'name attribution'.*

We agree that our findings to not directly address word learning necessarily, and have changed the title to “Young children's developing sensitivity to discourse continuity as a cue for inferring reference”

*5.      There does not appear to be a clear definition of 'discourse continuity'. It seemed that 'discourse continuity' and 'discourse structure' were used interchangeably. It became clearer after having read the experiments, but a definition earlier in the paper would be helpful.*

A definition of discourse continuity and distinction from discourse structure has been added.

*6.      The authors argue that the proximity of the utterances about the object to the naming event are insufficient, but that the inferences about the connections and continuity between the comments and naming presentations are essential. Specifically, the authors hypothesized that children would be more likely to infer an object's name if it is embedded between comments about the object as opposed to occurring after (or even before). However, it was not entirely clear how this hypothesis was generated from the literature reviewed in the introduction. The examples used as support of the hypothesis appeared contradictory, namely, that both proximity and continuity were helpful.*

We appreciate the reviewer’s comment. We believe that the literature reviewed in the introduction is still appropriate to support our hypotheses. Previous work has provided evidence that children can incorporate discourse context cues with social information to infer reference. Our experiments are the first to *isolate* the role of discourse structure in children’s reference disambiguation. No previous studies have dissociated the influence of proximity from continuity, therefore we review findings suggesting that both may play a role in children’s meaning learning, but our results from Experiment 2 suggest that proximity alone does not seem to be a sufficient strategy for children’s inferences about referential intentions, while discourse continuity alone appears to be a reliable cue for children where proximity alone is not.

*7.      The figures and graphs were sufficiently useful and clear (but see below regarding the need for further clarifications in the text). I was slightly confused by the proportion value in Table 3. Is this the mean of the proportion correct over the presentation orders*

Yes, and the proportion value description for Table 3 was clarified: “the mean proportion and standard error for participants' selecting the toy matching the naming location across trial types (i.e. choosing the first toy in a First Toy trial and the second toy in a Second Toy trial)”

*8.      The critical finding was that infants were more likely to select the correct referent at test when comments occurred before and after the naming event. What differed between Experiments 1 and 2 however, was not only whether the naming event was embedded between comments or not, but also that in Experiment 1, children heard one comment before the naming event, while in Experiment 2, children heard two. It is surprising that hearing two comments about a referent before a naming event has less effect than hearing one comment. The question that arose while reading this is what the effects would have been if both groups of children heard an equal number of comments about the object prior to the naming event, but one also heard an extra comment after. This way they would be matched on the events prior to the naming event, yet one group would experience the 'continuation' of comments. The groups would admittedly be unmatched on the total number of comments heard, but I feel more  
conditions are necessary to properly explore this question.*

We thank the reviewer for this idea. We don’t believe that hearing two comments before and one comment after should change how participants interpreted the discourse structure, but we ran this experiment on Amazon’s Mechanical Turk to see if this would influence performance for adults. We found that responding was identical whether one or two comments were provided before the naming event. Although we haven’t run this version with children, we expect that it would not dramatically alter their performance, and if anything may strengthen their sense of discourse continuity when more dialogue is devoted to a toy “topic”.

*9.      The design is not entirely clear. It appears from Table 1 that there are two presentation orderings. Did children take part in both or only one, in both experiments? I presume it is the former, given line 10 of page 10, but it would help if this were stated clearly. Are two different names used in the two presentations orderings? Table 1 implies that the same name was used for both orderings. It is also not specified in the text what the novel names were.  
10.     Further, if the children took part in the two presentation orderings, when were the children asked for the target? Was this after each ordering or after both?*

We clarified the design description to make it clearer that participants participated in a single trial per toy pair, and naming location was counterbalanced across participants: “In these Embedded trials, the experimenter introduced the naming event between two sentences about the same toy. The labels introduced in the four trials were "toma", "modi", "gazzer", and "zib". Each participant heard a naming event embedded between descriptions of the first toy for two of the trials (``First Toy'' trials) and between descriptions of the second toy for two of the trials (``Second Toy'' trials). Label location, trial order, toy pairs, and target side were counterbalanced across participants. “

*11.     Could you also state what the experimenter said when instructing the participants to select the target on page l0, line 10?*

We included the experimenter’s question to the procedure: “At the end of each trial, the experimenter prompted the participant to identify the named item by pointing (e.g. ``Can you point to the toma? Which one it the toma?").”

*12.     In Experiment 2, for the first presentation order it is inherently unclear to which object the name refers. The naming event is embedded between comments about both objects. When the name refers to the second object this is less likely to be confused. The authors acknowledge this difference on page 15, from line 18. However, in a sense the first order and the second order would have different effects and cannot be regarded as testing the same thing. They should therefore be analysed separately. The lack of difference in the results may be due to the fact that the two orderings were not comparable, as they were in Experiment 1.*

We thank the reviewer for this comment. If children were relying on temporal proximity, they should not necessarily have a bias for either toy when the name is presented between descriptions of each toy, but they should reliably select the second toy when the name appears after the second toy. However, when we individually compare each age group’s performance for each trial type against chance, no analysis is more than marginally significant. Although the two trial types had slightly different predictions, the lack of differences appears not to be due to comparisons across the groups, but rather because all groups results similarly around chance.

*13.     On line 41 of page 12, it would be clearer if the word 'second' could be placed between 'predicting' and 'toy'.*

Corrected.

*14.     The two experiments were essentially between group conditions of the same paradigm. There were no comparisons performed between the two experiments in relation to age and vocabulary level. This is necessary to rule out if children in Experiment 1 performed better because they were developmentally more advanced and not because of the type of discourse structure they were presented with. If any sort of vocabulary measure was obtained, a comparison between the two experiments would be of use.*

Due to the constraints of running this experiment at a museum, we were not able to obtain vocabulary assessments for our participants. We did collect some brief demographic information including parent level of education, and found no differences across groups.

*15.     Of question is whether there needs to be another condition as it is not entirely convincing that the naming event needs to be embedded between comments about the object. It is possible that in Experiment 2's 'After Second Toy' condition, after the children heard two comments about the unnamed object, then heard two comments about the next object, perhaps thought that the experimenter had completed making comments about that object and had moved onto something else. This is a remote possibility, but the two experiments are not matched in terms of the number of comments prior to the naming events, as well as differing in terms of whether the comments were embedded or not.*

We thank the reviewer for this comment. It is possible that participants could have expected that the speaker was moving on after the two comments in After Second Toy trials. If so, this would be consistent with our proposal that listeners use discourse continuity to make inferences about intended meaning, but lack a strategy when discourse structure cues are not available. Our Mechanical Turk results suggest that the number of comments preceding the naming event does not alter performance, so we believe it is not whether or not subjects heard two comments before the naming event, but rather whether the name is embedded between or stated after the descriptions.

*16.     The lack of comparability between Experiment 2's 'After First Toy' and 'After Second Toy' orderings (as mentioned in point 12 above) needs to be addressed. This issue precludes the possibility of comparing the effects of Experiments 1 and 2.*

Although we make different predictions about what patterns of results should arise given that participants rely on temporal proximity to make their selections, our analyses are still reliable in yielding these differences, if present. The only difference between Experiment 1 and Experiment 2 is the location of the naming event. We therefore included both experiments in a model to look for performance differences across the experiments. The results yield that participants are more likely to select the toy associated with the naming event only for the embedded trials and with increasing age. Because performance does not differ by trial type in Experiment 2, but reliably varies by trial type in Experiment 1, we are able to compare the overall results to demonstrate that the temporal cues alone do not seem to play a strong role in reference selection, while discourse cues play a much stronger role.

*17.     Given the questions outlined above, it appears that the evidence is not sufficient to warrant the authors' conclusion that the naming event needs to be flanked by descriptions of the object and that descriptions occurring in proximity to the naming event are not used by children to correctly infer a speaker's intended referent. More analyses with regards to the presentation orders or further testing are required.*

We appreciate the reviewer’s concerns. Because the only change from Experiment 1 to Experiment 2 is the location of the naming event, we believe that our design offers an informative control to support our findings that by age 3, children are able to use cues about discourse structure to infer reference. In Experiment 2, participants could have used temporal proximity to reliably select the toy closest to the naming event, namely the second toy in After Second Toy trials, but we found that this was not the case, and participants selected at chance in these trials. We believe this is evidence that participants in Experiment 1 were not relying on temporal associations alone, but rather were able to make inferences continuing topic cues within a discourse. We have run analyses indicating that no performance for either After First Toy or After Second Toy trials in Experiment 2 exceeds chance performance for any age group. Our experimental results also indicate that presenting one or two sentences before the naming event does not influence performance. Therefore, we believe our experimental results give sufficient support for our claims and an important contribution to a sparse literature on children’s comprehension of discourse.