Dear Dr. Baillargeon,

Thank you for the thoughtful comments on our manuscript, “Social cues modulate the representations underlying cross-situational learning.” Please accept our resubmission. We have addressed your comments and the comments of the reviewers, and we believe that the manuscript is substantially improved. Please find below a point-by-point response to the comments.

We’d like to particularly highlight our inclusion of a new experiment that shows an independent effect of referential cues even when inspection time is equivalent across the gaze conditions.

Please do not hesitate to contact us if you have any questions or concerns. We look forward to your consideration of this revision.

Sincerely,

Kyle MacDonald

**Editorial**

*Reviewer 1 raises an alternative interpretation of your findings, having to do with inspection time, that will need to be carefully examined, most likely through additional experimentation.*

We appreciate R1 suggesting this alternative explanation. We added the results of an additional experiment (Experiment 4) where inspection time is fixed across the Gaze and No-Gaze conditions. Replicating our previous findings, in this experiment we still found an effect of reduced memory for alternative word-object links in the presence of social cues. We also added analyses of inspection time to Experiments 1-3. We think both of these additions strengthen the manuscript.

*Reviewers 1 and 2 both raise the issue of whether a non-social cue (e.g., an arrow) would have the same impact as gaze on participants’ responses. It does not seem to me essential to address this issue experimentally in the present article, but it would be helpful to have some discussion of it in the General Discussion.*

We added a paragraph to the General Discussion that discusses this issue.

*Reviewers 2 and 3 both make suggestions for relevant prior work that should be included in the Introduction and General Discussion.*

We added the majority of the prior work that Reviewers 2 and 3 suggested.

*Like Reviewer 1, I had difficulty following parts of the design of your experiments (e.g., how did the interval manipulation work exactly?). Please describe one specific example to make this clear to readers, especially those not directly in your field.*

Thank you for pointing this out. To clarify this part of the design, we added the following sentence to the Design and Procedure section for Experiment 1. “For example, in the 0-interval condition, the test trial for that word would occur immediately following the exposure trial, but in the 3-interval condition, participants would see three additional exposure trials for other novel words before seeing the test trial for the initial word.”

*Your figures are somewhat difficult to follow, especially if printed in black and white. Is there anything than can be done to make them more reader-friendly?*

We made several changes in an effort to make the figures easier to interpret. Specifically, we (a) increased the font size of all text, (b) mapped the fill color for each point to the gaze and no-gaze conditions, and (c) mapped the type of line (dashed vs. solid) to Same and Switch test trials.

*As Reviewers 2 and 3 point out, there were multiple typos. Also, please follow APA style both in describing your research (e.g., use past tense) and in citing prior work in the text (e.g., & only in parentheses, no first initial except perhaps for K. Smith, and so on).*

We apologize for the typos. We thoroughly proofread the manuscript and fixed the citations to follow APA style.

**Reviewer #1**

*However, it’s not clear that the effect (less robust encoding of alternatives in the presence of gaze cues) is specific to social or referential cues. It is possible that any cue at all, for example a perceptual cue such as highlighting the main object (rather than gazing at it) could distract away from encoding the alternatives as robustly. A related possibility is that any concurrent task, regardless of whether it highlights an object, takes away memory resources and impairs encoding of alternatives. To claim that the presence of a referential cue selectively reduces participants’ memory for alternative word-object links, control conditions should rule out these alternative explanations.*

We agree with this interpretation in the sense that we think that our results would generalize to other cues that effectively communicate referential intent and/or modulate learners’ attention. With that said, previous research suggests that gaze cues result in a more reflexive attentional response compared to arrows (Friesen, Ristic, & Kingstone, 2004), that gaze-triggered attention leads to better learning compared to salience-triggered attention (e.g., via flashing objects) (Wu & Kirkham, 2010), and that toddlers readily use gaze to infer novel word meanings (Baldwin, 1993). As discussed above, we added a paragraph to the GD where we discuss the role of gaze as a cue, which we think clarifies our theoretical account.

*Experiment 3, which seemed promising, is somewhat confusing. To test whether participants respond to the grading of the reliability of gaze cues, why not manipulate the reliability of gaze cues directly, rather than manipulating reliability by changing the number of “switch” trials? The gazing informant isn’t unreliable per se in her behavior, it’s that more switch trials may give more evidence of alternatives-this seems unnecessarily indirect.*

We appreciate this point, but we think that increasing the ratio of Switch to Same trials actually is a relatively direct manipulation of reliability: Switch trials provide evidence that the speaker’s gaze is not a reliable cue to reference (because the gaze target does not show up in the subsequent test trial). This manipulation also has the added benefit of maintaining consistency between the familiarization block (where we establish reliability) and the test block (where we measure memory for alternative word-object links), which we think made the reliability manipulation less obvious to participants.

*Please discuss the findings in the context of the larger literature, including situating this work within the contradictory findings of Medina et al 2011 and Yurovsky et al 2013, making reference to a closely related paper by Yoshida et al, in Cognitive Science that compared statistical learning of words and tones and found differing memory strength for the alternatives to the high probability sound-object pair, and discussing the vast number of prior studies of social cues and word learning.*

We updated the general discussion to include more of a discussion about how our results fit within the literature.

*Please proofread. There are typos throughout the paper (e.g., in the abstract: undertainty ->uncertainty, on page 3 “with some parents’ rarely providing highly informative contexts and others’ doing so relatively often”, on page 21 intital -> initial, etc.).*

We apologize. We thoroughly proofread the manuscript to fix typos.

**Reviewer #2**

*I would like to see an additional experiment where inspection time is not participant-controlled (the objects would remain on screen for a fixed amount of time, independently of how much time the participant took to select one object), so that we can see whether the effects reported in the three experiments are entirely mediated by inspection time, or whether there is an independent effect of selective memory encoding in the presence of social cues;*

We appreciate this suggestion and have followed it in our revision. In particular, we conducted a follow-up experiment where inspection time was fixed across the Gaze and No-Gaze conditions. We found an effect of gaze, with worse performance on Switch trials after exposure trials where gaze was present. We think this result provides evidence that reduced inspection time does not completely explain the effect of social cues on learners’ reduced memory for alternative word-object links.

*In fact in Exp. 3 the authors could add another follow-up analysis to the ones they already have, by including a graph to show accuracy at test, as a function of how long they inspected the display during the corresponding exposure trial. Also, for panel B, would it not be as efficient to break it down by whether they selected the gaze-target or not on that specific trial (rather than by the number of trials in which a given subject followed the gaze)? In fact it would be best to present the two analyses together, one with whether or not they followed the gaze on that specific exposure trial, and one with how long they inspected the display.* *You could perhaps even consider running an analysis that would pit one factor over the other, if they are sufficiently decorrelated in participants' behavior. This would not however remove the need for an experiment in which inspection time is decorrelated from presence/absence of the gaze cue.*

Thanks for the idea. To address the questions about the role of inspection time, we added analyses of this variable to Experiments 1-3. However, when we modeled the length of inspection time, participants’ use of the gaze cue was was the strongest predictor: when gaze was present and participants used it, inspection times tended to be shorter. This confound made it difficult to know whether the effect of gaze on Switch trial performance was mediated by a reduction in inspection time (as you pointed out). We think that the addition of the fixed inspection time experiment helps to address this issue.

We also changed the figure in Experiment 3, which we think makes the results clearer. We split the plots into two figures: Figures 4 and 5. Figure 4 shows (a) the primary analysis of accuracy as a function of reliability and (b) a new plot, showing accuracy as a function of reliability and whether participants followed gaze on exposure. This plot shows that participants were more accurate on Switch trials when they chose not follow gaze and that this effect was larger when the gaze cue was less reliable. Figure 5 shows the secondary analyses of (a) the number of trials participants chose to follow gaze and (b) subjective reliability judgments. We chose to keep the “gaze use” analysis shown in Panel A of Figure 5 since this was a confirmatory analysis and provides evidence that the effect of reliability also occurred at the participant level.

*The examples provided illustrate only one condition though, I wondered for instance how you displayed the 6 or 8 objects, perhaps if you added more experimental conditions to your on-line demo, this would make things clear to readers without making the description of the experiment too heavy.*

Thank you for pointing this out. We updated the project site to include all experimental conditions (0-, 1-, 3-, and 7-interval; and 2-, 4-, 6-, and 8-object) for Experiment 1. We also added the fixed inspection time experiment (Experiment 4).

*Do you think that a pointed arrow would have the same impact on participants than gaze? would you count such a cue as ‘social’ or not? Depending on your answer to these questions, you may want to rephrase somewhat your discussion of the gaze cue as a specifically social cue, and reduce the dichotomy between ‘statistical’ and ‘social’ cues that you currently set up, and think instead of a multidimensional space where cues as to the possible meaning of a word may come from many directions.*

We do think our results would generalize to other cues that effectively communicate referential intent and modulate attention. With that said, previous research suggests that gaze cues result in a more reflexive attentional response compared to arrows (Friesen, Ristic, & Kingstone, 2004), that gaze-triggered attention leads to better learning compared to salience-triggered attention (e.g., via flashing objects) (Wu & Kirkham, 2010), and that toddlers readily use gaze to infer novel word meanings (Baldwin, 1993). We added a paragraph to the GD where we discuss whether gaze is a privileged cue to social attention, which we think clarifies our theoretical account. We chose to maintain the structure of the paper and to interpret our results within a framework of how social cues constrain the input to statistical learning mechanisms.

*You discuss participants’ responses on exposure trials (for Gaze trials, e.g. on p.14, p. 18), please systematically give %correct rather than only giving the stats (sometimes these values appear later on, but in the discussion, so it is a bit confusing).*

We added the overall mean for gaze following on exposure trials (% correct) for each experiment. Since there were 32 conditions and more variability in this measure in Experiment 1, we also included the range of % correct across all conditions.

*Also, the definition of what is a ‘correct answer’ varies, so please make it very explicit – when judging a test trial, a correct answer is the object chosen by that participant on the exposure trial, for a same test trial, and another object present in the exposure trial, for a switch test trial; but for exposure trials with gaze it is the object that was gazed at – for same test trials readers can be confused as to what is the correct response (the object that was gazed at, vs the object that was selected by the participant).*

We added the following sentences to the analysis plan: “Correct behavior on Exposure trials was defined as selecting the referent that was the target of gaze in the Gaze condition (note that there was no "correct" behavior for Exposure trials in the No-Gaze condition). Correct behavior on Test trials (both Same and Switch) was defined as selecting the referent that was present during the Exposure trial for that word.”

*p.10 last paragraph: I agree in principle with not including multiple-way interactions that are difficult to interpret. However, the fact that you include a 3-way interaction in your analysis of the exposure phase but not in the test phase; and that you argue with the lack of hypothesis for the 3-way interaction and with a lack of increased model fit for the 4-way interaction, seems a bit arbitrary. You could perhaps base your criterion for the exclusion or inclusion of an interaction on the same criteria for all analyses.*

Thank you for pointing this out. We now limit all of the models in the paper to include only two-way interactions. To reflect this change, we added the following sentence to the analysis plan: “We limited all models to include only two-way interactions because the critical test of our hypothesis was the interaction between Gaze condition and Trial Type, and we did not have any theoretical predictions for possible three-way or four-way interactions.”

*p.18 “The interaction between speaker reliability and subjective reliability assessments was marginally significant (b = -4.33, p = 0.1).” This is not even marginally significant. You could point out the descriptive tendency instead.*

Thank you for pointing this out. We have changed our reporting of “marginal” p-values in the main text. We now only interpret effects that achieved statistical significance at the .05 level, but provide the full model output for each model in an Appendix.

*fig.3:: initially I saw only one line for the B graph, ‘same’ trials, and was very confused (I had to blow up the pdf quite a lot to see the two lines). Could you do something to improve this? Otherwise mention it in the figure caption.*

We jittered the points and made them larger in order to make it clear that there are two lines in the plot.

*-p. 17-18: ‘Since we were no longer testing the effect of the presence or absence of a referential cue, all exposure trials in Experiment 3 included gaze, but this cue was more or less reliable depending on which familiarization block participants saw.” this sentence is ambiguous: I assume that what you did was have an equal number of Same and Switch trials, at test, as you did before; however said that way it looks as if you also manipulated gaze reliability at test, with more or less Switch trials depending on the condition.*

We removed the final clause in the sentence, which we think makes it less ambiguous. It now reads, “Since we were no longer testing the effect of the presence or absence of a referential cue, all exposure trials throughout the experiment block included a gaze cue.”

*p. 18: offset term: I assumed you always had 4 targets, as in Exp. 2; why do you need the ‘offset’ term then?*

The offset term encodes the chance probability of success given the number of referents in the task. So even though the number of referents does not vary in Experiments 2-4, we still included it.

*p. 21 typo intital instead of initial*

Fixed.

*p.7 3rd paragraph: “test trials occurred after either 0, 1, 3, or 7 trials”, please clarify this, it is difficult to understand unless one goes on-line to check the experiment.*

We added the following example to the method section, “For example, in the 0-interval condition the test trial for that word would occur immediately following the exposure trial, but in the 3-interval condition participants saw three other exposure trials before seeing the test trial for that word.”

*p.7 3rd paragraph: “At test, gaze was never informative”. Change to “At test, gaze was always directed straight ahead”.*

Fixed.

*p.15 last paragraph: Please mention how many participants were not following gaze reliably in Experiment 2.*

Fixed.

**Reviewer #3**

*First, I think the authors need to expand their discussion of the limitations of their study, which in turn will require further nuances to their theoretical account. In particular, I think they should comment on their requirement that subjects were forced to select a referent on every word-learning instance. Although there is some evidence that overt selection of referents doesn’t significantly alter findings in experiments like this, it may in this case. In general, the experimental setting strongly suggests that each word’s utterance is intended to refer to a co-present object. But we know that much of language doesn’t work this way. The presentation of a miniature world and a set of objects to pick from (as one does here) suggests to the learner that all words can be learned from observation of the referent world. If instead a learner is willing to entertain the possibility that speech isn’t about co-present objects, then especially when speaker-attention cues are absent, learners might assume this option (that the speaker is talking about other things not immediately present), and learn nothing about the word’s meaning, thereby making no word-to-world link – opposite of what was found here: namely that learners try to learn even more from the co-present referent world when social cues to reference were removed. I don’t think this is a game-changer, but it seems wise to include such a discussion given the reality that most adult language use isn’t about the here-and-now, even speech to kids is often about things not co-present (as Gillette et al., 1999, showed rather nicely).*

We appreciate the point about the gap between our task and real-world learning contexts. We expanded several aspects of the limitations section of the GD to reflect these challenges. Specific to the current point, we include a discussion of how features of our task may have biased participants to assume words referred to a physically co-present object. And that when language does not refer to physically co-present objects, learners would not benefit from storing additional word-object links in the absence of referential cues.

*Second, when talking about whether similar findings might obtain in children, the authors might wish to refer to the work of Woodard, Gleitman & Trueswell (in press, Language, Learning & Development) who use the same-switch paradigm and find that young children show little or no signs of retaining unselected referential alternatives. This can in principle be seen as consistent the with the authors present account. Second, Trueswell et al. (2016, Cognition) examines the social cues to reference in natural parent-child interactions, their frequency and types. This paper too seems highly relevant to the current discussions. Although this sounds like a self-centered request from one of the authors (okay, it is), I believe that the work is also truly quite relevant. I suspect those findings will only enhance the author’s discussion and account.*

Thank you for pointing us to this work! We added both articles to our discussion of how the current results would generalize to younger language learners and we think the discussion is stronger because of it.

*Third, it seems like something went wrong in the following sentences on page 3:*

*“Using the HSP, Medina, Snedeker, Trueswell, & Gleitman (2011) found that adults did not aggregate multiple word–referent correspondences across trials, concluding that real world learning contexts are too noisy to support tracking of multiple word-object links. In contrast, Yurovsky, Smith, & Yu (2013) found a bimodal distribution, with half of the naming episodes being unambiguous to adults and half being quite clear.” This quote doesn’t describe the Yurovsky et al. (2013) findings correctly. Figure 1 in the Yurovsky, Smith & Yu (2013) article shows a histogram, providing the proportion of naming episodes that fell into different levels of accurate guessing by subjects (HSP accuracy from 0% correct to 100% correct).*

*It is clear from that figure that only 10 percent of the naming episodes were truly unambiguous to adults (that is, the right most bar is 100% correct guessing of the intended referent, and is around 0.1 proportion of all naming episodes). I suppose if the authors want to include the 90-100% correct as unambiguous then that would be about 30% unambiguous. Either way, clearly it is not the case that “half of the naming episodes” are “unambiguous to adults”. The authors also say that the other half were “quite clear”. Did they mean quite unclear? The other bump in this bimodal distribution includes 0-25% HSP accuracy scores (the left bump). Surely these are unclear situations. This should be corrected and clarified. The second concern about the above quote is the direct comparison between these findings from Yurovsky, Smith & Yu and the findings from Medina et al. (2011). Overall average HSP accuracy from Yu et al. was 58% correct, which is much higher than all other reported HSP studies in the literature. Trueswell et al. (2016, Cognition) reports that HSP accuracy was about 17% from Medina et al. (2011) and 22% from Cartmill et al. and 28% in Cartmill et al. But, there is a good reason why Yurovsky et al. (2013) got much higher scores (more than twice as high). Yurovsky et al. (2013) only used videos of parents engaged in object play, and critically only sampled utterances for which the parent labeled a co-present object of interest. That is, if the mother said “giraffe”, that was selected as an HSP video because the toy giraffe was one of the objects they were playing with and was an object of interest. But if the mother in that same situation uttered “remember the doggie we saw yesterday?” That instance would not be allowed to enter into their HSP study, because the parent and child were not playing with a toy dog, and the dog was not an object of interest in the study. This is connected to my concern above: If one is going to use HSP accuracy scores as a measure of the degree of referential ambiguity, surely estimates should come from randomly selected examples of mothers uttering concrete nouns of all kinds, not just the cases in which the object in question is already present. This latter strategy was used in the other HSP studies mentioned above, and not surprisingly HSP accuracy is much lower.*

Thank you for pointing out. We fixed the typo, and we made our description of the Medina et al. (2011) and Yurovsky et al. (2013) findings more precise. We also added a footnote where we discuss how the different sampling procedures could lead to different estimates of ambiguity in the input, and we make it clear that our goal was to point out that variability in referential uncertainty exists and thus could affect statistical learning mechanisms.