**Replies to reviewers’ comments:**

**Reviewer 1**

*This is an interesting paper that presents valuable new data on the acquisition of the capability to infer ad hoc implicatures. My main concern with the paper as it stands is that it doesn't quite do justice to the theoretical issues around implicature, particularly with respect to the question of whether the participants' choices are ultimately motivated by implicature in the strict sense, or some less categorical inference.*

*This point was already made in the previous round of review, but just to underline it: consider the discussion of Stiller et al. (2015) on p.5. There, children are asked to choose between a face with glasses and one with glasses and a hat, based on the description "My friend has glasses". This certainly could be an ad hoc implicature (with "glasses" interpreted as "glasses but no hat"); indeed, it could even be an exhausitivity inference (with "glasses" interpreted as "glasses but no other distinguishing features"). It could also be a probabilistic best guess - suppose we're equally likely to make reference to the glasses or the hat, and we choose one feature, and we're equally likely to be talking about each face. Then two-thirds of the time we refer to "glasses", we're talking about the face with glasses and no hat, and hence that's the one a rational hearer should (always) choose, as their best guess. (This is perhaps implicit in the turn of phrase the authors later use on that same page, where they talk more carefully about "the implicature-consistent choice", rather than, say, "the implicature-driven choice").*

*This may seem like a rather pedantic objection to raise - who cares whether these inferences are strictly speaking implicatures or not? However, as it stands, the key point of this paper is to argue that (specifically) implicature and salience are competing factors in children's decision-making in this kind of experimental paradigm. It strikes me quite forcibly that rational non-implicature inference could also be in the mix, and would compete with salience too. Moreover, the variation in the number of features of the competitor presents a very natural way to explore precisely this question, so I think it should be discussed.*

The reviewer raises the important point that there are many accounts of implicature present in the theoretical literature. These accounts include standard Gricean and neo-Gricean accounts, grammatical accounts that use an EXH operator to exhaustify possible alternatives, and a variety of game-theoretic and probabilistic accounts that use recursive reasoning (listeners reasoning about speakers and vice versa) to explain Gricean behaviors. This theoretical space is complex, and the current study does not allow us to differentiate between these proposals.

That is not to say that we do not have a position on this topic. Goodman & Frank, 2016, *Trends in Cognitive Sciences* reviews the rational speech act framework for probabilistic computation of pragmatic inferences; Potts et al. (2017), *Journal of Semantics* gives some thoughts about how grammatical and probabilistic views can be integrated. But comparing these viewpoints using data is complex, as the predictions of each model are not always well-specified. For example, we believe it’s tricky to differentiate what the reviewer calls an “ad-hoc” vs. “exhaustivity” inference here – in both, some alternatives are negated, but the question for all theories is which alternatives those are! No current theory gives a good account of how such alternatives are specified in these kind of situation-specific inferences.

For the purposes of the current manuscript, however, one point is important. We believe that all of the alternatives on the table are implicatures, thus the claim that we are studying implicature is well motivated. The competitor suggested by the reviewer is a “probabilistic best guess” account. On close inspection, this account looks almost identical to a rational speech act model of the type described in Goodman & Frank (2016) and Frank & Goodman (2012): a rational listener reasons about a speaker’s choices given particular communicative aims, then makes a “guess” based on those probabilities. Those models are intended to capture implicature behavior. The idea of a guess – a probabilistic choice – is inherent in linking any probabilistic model to human responses; it should not be construed as differentiating the behavior from what is conventionally meant by the term “implicature.” (Whether RSA models capture implicature by making it more continuous with general ambiguity resolution – and hence remove the need for a separate term – is a deeper philosophical issue, of course!)

Based on the reviewer’s comment, we have rewritten nearly the entire introduction. While it is longer, we hope the reviewer now will agree that it describes the theoretical space more extensively and more accurately. This rewrite also gave us the opportunity to discuss the relationship of our current experiments to the RSA model.

*There are several other areas where I felt that the paper could be improved, generally because of slightly loose appeals to theory: I note these below.*

We apologize for the theoretical looseness – as noted above, we have tried to avoid making unwarranted theoretical claims, but we do believe that our work here should constrain theorizing about implicature development. We have tried to clarify the direct connections to the RSA model as well as caveating some of the other theoretical claims.

*Lines 81-83: "In this task, there was no need for children to spontaneously generate the alternative ("glasses and hat")...because the alternative was visible in the context." That isn't really true: "the alternative" is a linguistic expression, and what is visible in the context are some objects. So there seems to be quite a strong assumption about preferred encoding lurking within this remark.*

Thanks, we now have rewritten that sentence. The revised text reads:

“In this task, the alternative referent (face with glasses and hat) was visible in the context, and thus access to the alternative terms ("glasses” and “hat") was made easier. In general, we assume that the standard route for referring to these visual properties of the context will be by naming them. The design intention in this study for using simple nouns like “hat” was therefore to make it obvious what the linguistic alternatives would be by virtue of the highly accessible names for stimuli.”

*Lines 112-114: "Further...what referent a speaker is talking about." I found this point a bit elusive: could it be made more concrete?*

Thanks, we have now expanded this point with a more concrete example and explanation.

*p.7, footnote 2: Quite a bit of argumentation and a couple of conclusions seem to be brought into this footnote - I think they'd be better later in the main text.*

Thank you, we have now moved the content of this footnote to later in the text, in the Discussion section.

*Line 205: "due to the strengthening of implicatures". Could you be clearer what you mean by this?*

Yes, thanks, we now have modified the wording of this part based on feedback from all three reviewers:

“We hypothesized that older children would choose the target more often in the more-feature implicature trials than the fewer-feature implicature trials due to the strengthening of implicatures – "Elmo's lunchbox has an apple" is more likely to mean "apple only" given an orange AND cookie on the alternative referent, thus more things that could have been named but were not. On the contrary, younger children were predicted to choose the target less often in the more-feature trials than the fewer-feature trials due to increased saliency of the distractor.”

*Lines 274-276: "indicating that the speed of implicature computation did not improve with age as much as the speed of processing unambiguous meanings". I'm not clear on how your analysis distinguishes between implicature computation and the guessing of implicature-compatible answers (let alone the other possibilities I'm arguing for). It seems a bit premature to conclude anything about implicature computation based on data that must uncontroversially include some points that don't involve implicature and were contributed by participants who were not competent with implicature. (Presumably, if three-year-olds are getting, say, 70% rates of implicature-compatible responses, in a binary choice task, that would suggest that crudely speaking 40% of them are figuring it out correctly and 60% of them are guessing. In that case, 3/7 of the correct responses are just guesses, the timing of which doesn't say anything about implicature on anyone's analysis.)*

As a starting point to this comment, we want to note that our interpretation of reaction time data is very limited. We write that “speed of implicature computation did not improve with age as much as the speed of processing unambiguous meanings.” We now revise this to read “reaction time on implicature trials” rather than “implicature computation.”

More broadly, we agree that choice reaction time is complex to interpret; in the case of true uncertainty about whether the subject has a particular piece of knowledge, it’s not correct just to say that the mean RT is the speed of the computation. That said, we do not think that the two are unconnected, and we hope we can clarify our reasoning.

First, there are two different “guessing” mechanisms being discussed. The first is the “probabilistic guess” account described above. In our response, we hope we have convinced the reviewer that such an account is in fact an account of implicature under a model like the Rational Speech Act model; indeed, this is the account we favor. The second kind of guess, which we will treat here, is a “random guess.” This kind of guess reflects an uninformed choice – think, pressing a side randomly, without listening – rather than a probabilistic judgment on the basis of some uncertainty. In any psychological task, there will be some proportion of trials on which participants guess randomly (even if this proportion is very low, for example, with adults in an easy task).

But in fact, most psychologists don’t think that “random guessing” is why participants get trials wrong. Instead,when participants get the “wrong answer” on a task, they are trying to answer correctly but failing. Modern decision-theoretic models, like the drift diffusion model (e.g., Ratcliff et al., 1999), quantify the tradeoff between speedy responding on a task and accurate responding – the faster you do a complex task, the more you will fail even though you are not trying to guess. (Then there is some small proportion of cases where people are actually failing to do the task; some models add a parameter to deal with this). Although in some of our lab’s work we have attempted to use these models with children’s data on implicature tasks (e.g., Schneider et al., 2016), this approach raises a complex set of issues and we don’t believe it’s appropriate here. Fitting such models requires a lot of data and is really at or beyond the state of the art for developmental data; there are only a small handful of papers using DDM with children.

So what can we conclude by looking at reaction time data in this task? We can see the average time to reach a decision – correct or incorrect. If children are *both* faster *and* more accurate in a task (as happens with development in familiar word recognition in our study and many others), it seems clear that they are getting better at that task and there is no speed-accuracy tradeoff. For our implicature condition, we also see that they are getting faster and more accurate, just at a slower rate. If we were seeing “fast guessers” becoming “slow implicature computers” we would predict a speed accuracy tradeoff, which we don’t see.

In sum, we hope the reviewer agrees that our findings are consistent with the modest claim we make, that is: developmental improvement in the implicature trials.

*Lines 279-291: I think the attempt to understand the processes undertaken by the participants is welcome, but the discussion in this paragraph struck me as a bit too speculative.*

Thanks, we agree with your opinion – to make sure it is clearly conveyed to the reader that this possibility is only speculative, we now have added: “The potential advantage of identifying a feature when it is by itself is only speculative, however, and should be examined further in future work.”

**Reviewer 2**

*I recommend that this paper be published with minor revisions. (Note, although this is a new manuscript, I have commented on these experiments before when they were a part of a larger paper. Most of the concerns I had about the previous draft with respect to these experiments have been addressed in this new, shorter version.) There are only two contentful comments I have.*

Thanks, we are glad we were able to address most of your concerns for the previous manuscript. We are grateful for the reviewer’s thoughtful commentary that led to the current, improved version.

*A) The authors should not talk about informational strength with respect to lexical items. Only propositions convey information and both "some" and "all" (for example) can be used in both stronger or weaker propositions. So, for example on line 33, the authors say "all, which is more informative than the alternative some”. "Some" is not more informative than "all". Information is communicated by propositions, not words. "John ate all of the cookies" is more informative than "John ate some of the cookies" but "Everyone who ate some of the candy in their hallowe'en bag was sick" is more informative than "Everyone who ate all of the candy in their hallowe'en bag was sick". Sometimes "all" is used in more informative propositions and sometimes in less informative propositions. There is no inherent information value to the word independent of its use to communicate a proposition. The same mistake is made on line 323.*

Thanks, we have now fixed these errors.

*B) On lines 203 to 207, the authors write "We hypothesized that older children would choose the target more often in the more-feature implicature trials than the fewer-feature implicature trials due to the strengthening of implicatures, and that, on the contrary, younger children would choose the target less often in the more-feature trials than the fewer-feature trials due to increased saliency of the distractor." I think this is awkwardly worded and, at best, ambiguous. It makes it sound like the authors predict that the older children will have the opposite influence in terms of object salience (i.e., that the more salient the non-target is, the less likely it will be chosen). I don't think this is what the authors intend to say. (If it is, then I don't see how this prediction would follow.) What I think the authors intend to say is that older children will behave in a way that is more consistent with computing the strengthened meaning, whereas younger children will be more likely to be distracted by choosing the object that is more salient but still consistent with the literal meaning. These are what the results demonstrate. I think these lines should be re-written to more accurately reflect the authors' prediction.*

Thanks very much for this helpful point, we have now revised the wording, as discussed in our response to R1.

*C) In lines 317 to 324, the authors suggest that their salience hypothesis could extend to account for children's behaviour in other tasks, such as felicity judgment tasks and truth value judgment tasks. They claim that it remains "an open empirical question whether the salience mismatch account might explain children's difficulty with these other cases of implicature as well." However, given the nature of the explanation, I don't see how this is possible. In the current task, the children have to choose between two pictures. In the critical trial, one picture is less salient but matches the pragmatically enriched meaning of the sentence whereas the other is more salient but only matches the literal meaning of the sentence. The idea is that children get distracted by the more salient object and choose it instead of choosing the less salient one that matches the enriched meaning. In the other tasks, there is no choice between objects. There is just one situation and children have to say whether the sentence is true/silly in that situation, or they have to say whether the speaker who uttered the sentence is right/silly. There is no competition between two different scenarios. Object salience is not a possible complication. The authors suggest that "salience" factors might apply to the choice of interpretation (enriched versus not enriched) or the "salience" of alternatives, however no plausible explanation is given of how this could be a factor. Either the authors need to spell out a plausible influence in detail or they need to admit that object-salience is a confounding factor in this task only. I suspect the later option is the only feasible one. I cannot think of how object salience could play any critical role in the other types of tasks that are used to assess children's abilities to compute implicatures.*

Thanks for this important point to consider. While we do agree that not all tasks of implicature computation involve referent selection and thus there can be less influence of salience of alternatives as we address here, we do think that the salience bias may play an important role in other experimental tasks that use children’s judgments between possible answers with different degrees of perceptual (visual) or conceptual bias. Our revision highlights the way that the RSA model analyzes perceptual salience as contributing to a prior probability distribution over referents (and hence suggests the analogy to other pragmatic tasks).

Outside of this theoretical connection, there are some further reasons to think that this account may apply more broadly. First, there are other works that have looked at a similar object reference to look at children’s processing of scalar implicature (e.g., Horowitz, Schneider & Frank, 2018) where you see some evidence of similar biases toward more salient alternatives (e.g., a picture where “all of the pictures are cats” as opposed to “some”). Second, there also can be competition between verbal answers they could give (e.g., true/right vs. silly). There are some previous works that talk about potential biases children have toward one verbal answer compared to another (though not in implicature literature -- see below, for yes bias to yes-no questions). Indeed, we see similar biases toward incorrect “yes” responses (compared to fewer incorrect “no”s) in Papafragou and Musolino (2003) and Barner et al. (2011) even in conditions where children computed literal/pragmatic meanings relatively well. We now have added a paragraph with this last point.

*D) The statement in lines 377 to 379 is too strong. It states that "that younger children’s failures with implicatures are likely related to effects of the salience mismatch between possible referents". I think this hold for the specific task that the article is talking about but it cannot hold for children's failures with respect to other types of tasks. The authors should temper this claim by saying that "younger children's failures with implicatures on an referent-choosing task are confounded by the salience mismatch between possible referents". This is the conclusion that the evidence from their experiments supports. Anything stronger is beyond the scope of their paper.*

We agree that the generic, present tense wording of this sentence was too strong. We’ve refocused the sentence on our current findings and the interpretation of this particular context manipulation.

**Reviewer 3**

*This is an interesting study on children’s ability to compute ad hoc scalar implicatures between the ages of 2 and 5. I have seen an earlier version of this work and the current manuscript is empirically richer and more focused. The study itself is wonderfully done and clearly presented. It is also theoretically more precise, even though some concerns remain.*

Thanks, we are happy to know that the current manuscript is improved from the previous version.

*The main contribution of the paper continues to be a replication of prior findings on children’s understanding of ad hoc implicatures, and evidence for the idea that implicature calculation drops when the distractor is more ‘salient’ (i.e., visually busier). This finding has some value in a rapidly developing literature on pragmatics but is still limited in at least two ways: (a) It only applies to methodologies that involve a forced choice between alternative visual scenes, and (b) it relies on the assumption that choice of scene means implicature calculation (and not, for instance, the likelihood of choosing a word such as ‘apple’ to refer to each of the two scenes in Fig. 1). Both of these concerns were raised during the review of the previous version of the paper. The present version has tried to address the first concern (see Introduction) but the text remains vague about how results can be generalized to other paradigms. (In the Conclusions, the paper attempts a connection to the mutual exclusivity literature that has also used visually contrasting scenes as stimuli. But I found the connection very indirect – there are multiple other differences between assigning novel word meanings and calculating implicatures that may explain early success with exclusivity that do not characterize implicature derivation.)*

Thanks for these thoughtful points. With respect to the first, in our revision we have attempted to be as clear as we can about what we see as the connection between the current paradigm and other paradigms. We have to be up front, however, about the fact that this connection is speculative and based partially in our particular theoretical framework (the Rational Speech Act model, see our response to Reviewer 1). Therefore, we have undertaken a fairly significant rewrite of the introduction to highlight these connections.

Despite this, we emphasize that reference resolution is in fact an important – perhaps even the default – mode of language use for young children. The use of pragmatic inferences to resolve uncertainty between referents is likely very common, whereas the frequency of true scalar implicatures is unknown in child-directed speech (and it is known to be relatively low even in adult-directed speech; Degen, 2015, *Semantics and Pragmatics*).

A further benefit of the salience account is that it raises important methodological considerations, not only for referent-selection paradigms (with *visually* salient distractors) but any paradigms where participants might have biases toward more *conceptually* salient answers.  
Bias due to salience seems to exist even in verbal answers to questions, e.g., younger children (2-year-olds) show a bias toward yes compared to no in answering verbal questions but this bias goes away with age (Fritzley & Lee, 2003), and there is some evidence that this bias is related to both their verbal ability and inhibitory control (Morigushi, Okanda, & Itakura, 2008). Thus this bias toward perceptually or conceptually salient answers need to be taken into account in designing tasks for younger children, not only for implicature computation but for any domains that make use of children's judgments that are potentially biased between possible answers. We have now added these points to Discussion.

*The second concern about the linking assumptions between theory and tablet data has not been addressed in this version.*

As we discussed in our response to Reviewer 1, we view choice behavior as reflecting some underlying preference for the implicature-consistent referent. While our task does not differentiate between different varieties of implicature, we believe that alternative accounts all fall into the category of alternative implicature accounts. For example, we do not understand how the reviewer intends that children compute “the likelihood of choosing a word such as ‘apple’ to refer to each of the two scenes in Fig. 1” without invoking some variant of a Rational Speech Act model, a probabilistic model of pragmatic inference. We hope that the rewrite of the introduction gives a much more thorough treatment of these issues.

*Furthermore, the paper could do a better job of stating and motivating its hypotheses. The introduction does not state clearly what the main goal of the paper is (see p.7), or the predictions. When predictions do appear, they are not motivated in their entirety. For instance, it is hypothesized (p.10) that “older children would choose the target more often in the more-feature implicature trials than the fewer-feature implicature trials due to the strengthening of implicatures, and that, on the contrary, younger children would choose the target less often in the more-feature trials than the fewer-feature trials due to increased saliency of the distractor.” The origin of the first hypothesis is unclear: what is ‘strengthening of implicatures’? Is this simply an intuition or is the hypothesis based on a theoretical assumption? Only the prediction concerning younger children has been mentioned so far so more justification of the prediction concerning older learners is needed. Notice that the data do not offer evidence for this last hypothesis (even though this may be due to a ceiling effect).*

Thanks, we agree with this point. In our rewrite of the introduction, we have introduced our predictions much sooner and in much more depth. We think a lot of the issue here came from the fact that we were reasoning from the Rational Speech Act framework but did not clearly say so.

*Minor points:*

*Fn.1: “For our purposes, this entailment relation is still ad-hoc in the sense that “the cookies” does not entail “the chocolate chip cookies” in discourse contexts in which no chocolate chip cookies are part of common ground”: the notion of entailment is independent of discourse context.*

Thanks for pointing this out. We now changed the wording to: “For our purposes, this relation is still ad-hoc in the sense that there is no reason for “the cookies” to implicate “the chocolate chip cookies” in discourse contexts in which no chocolate chip cookies are part of common ground.”

*P.4: “adults robustly compute implicatures, albeit sometimes more slowly than unambiguous literal meanings”: the second claim here is actually debated (see more recent papers by Tanenhaus and Grodner and response from Huang and Snedeker in Cognition) so the text needs to be more nuanced.*

Thanks, we now say: “... though their processing time can vary depending on the context”, and we have added the references you mention.

*p.15 and earlier: “by 2 years of age, children begin to be aware that informativeness is important to communication”. I agree with the summary of the cited papers – but there is also a long line of papers reporting massive failures to provide informative descriptions in referential communication tasks (unless children are asked for clarification or given feedback after an initial unsuccessful attempt; see work by Matthews, Tomasello, Katsos etc.). It’s a puzzle why pointing leads to early successes but verbal communication is often not as informationally rich.*

Thanks very much for this point. We think this point about young children’s difficulty in producing informative utterances actually underscores children’s challenge to have access to relevant information in comprehension and production -- both speakers’ need to *be* more informative, and exactly *how* speakers can be more informative. Our study makes both of these things accessible in the context. We now remark on this issue in the fourth paragraph of the Discussion section, as this issue is relevant to implications from both Barner et al. (2011) and our study.