Editorial Comments

> *Reviewer 1 responded to the point you made about how it's not clear whether the effect of informativeness is consistent across your subjects. I was impressed by the fact that the authors even raised this point; it's something that should trouble us researchers all the time, but we generally ignore it. You clearly can't address the problem in Experiment 1, and I doubt that four trials is enough to let you address it seriously in the other experiments, but perhaps you could see if there is anything in the distributions of choices that is informative.*

We thank the Editor for this comment. Please see our response to Reviewer 1 for details on our reanalysis. In brief, we find no evidence for idiosyncrasy in the distribution of children's responses, leading us to maintain the null hypothesis that informativeness is consistent across subjects. Admittedly, none of our studies were designed to examine this question, however, and in other unpublished work we have found that designs that repeat these sorts of inferences too many times lead to learning over the course of the experiment.

> *The reviewer also expresses a concern about the real-life applicability of your findings. At the very least, you should make it clear what the photos showed in Experiments 2 and 3, and discuss the limitations. I think that the reviewer, and I, would be much happier if you could replicate the last experiment with pictures of real objects, staged of course with salient contrasting attributes.*

First, in response to this comment we have added an Appendix with the full stimulus set. More broadly, we also see two interpretations of Reviewer 1's comment, as being about either 1) feature salience, or 2) the difference between pictures or schematic stimuli vs. real-world objects.

The first point, about feature salience, we view as extremely important. As we noted at the end of the Model section, features in the world vary not only in their informativeness but also in their salience. A feature may be informative but subtle, or highly salient but uninformative. This distinction is important, and in our previous work on this topic (Frank & Goodman, 2012), we provide a quantitative account of the role of salience in pragmatic computations. Salience is clearly an important topic for children's word learning. In our revision we have added a "salience control" in Experiment 4 showing that pure perceptual salience judgments do not explain our findings.

On the other hand, we would like to argue against the importance of the second point regarding differences between illustrations, pictures, and 3D objects. From earliest infancy, babies respond to pictures and schematic displays in the same ways that they respond to the real thing (e.g. M. H. Johnson et al., 1991 for face perception; S. P. Johnson & Aslin, 1995 for objects), and—because of this—nearly every experiment in the developmental literature relies on such schematic displays. In the word learning literature, there is a systematic study of the differences between objects and drawings by Preissler & Carey (2005), showing convincingly that even 18-month-olds (much younger than our participants) treat schematic line drawings identically to real 3D objects for purposes of word learning.

Thus, while we have added critical experimental evidence on the role that salience plays in our findings, we do not believe a replication with photos is warranted here.

> *Reviewer 2, in addition to the concern about the appropriateness of the paper for Cognitive Psychology, makes a large number of valuable suggestions. Dealing with all of them would certainly improve the paper. Frankly, I don't know how I would deal with the apparently-real deviations from the model predictions in Experiment 1; I hope you can think of a way.*

We have modified the manuscript to address some of these comments, reviewed below. We also address the question of apparent deviations from the model below. In brief, we think that the important result is that the model has a very high linear correspondence with the human data. It is common practice in psychophysics and other disciplines to assume that a model should not predict 0% and 100% performance even in completely unambiguous situations, but should instead allow for "lapses" by the human participants. Such an adjustment leads to perfect fit (within the measurement error of our experiment and the—admittedly limited—number of measurements we made). See below for more discussion.

> *Reviewer 3 raises a very different concern: is the essence of your findings what they say about word learning, or about how your subjects viewed the objects? I'm not certain that the experiments the reviewer proposes are quite what you'd want to do, but some discussion is certainly needed. I'd wonder, for instance, whether the distinction between informativeness affecting 'pragmatics of word learning' and affecting 'shared attention' is one you'd want to put much weight on.*

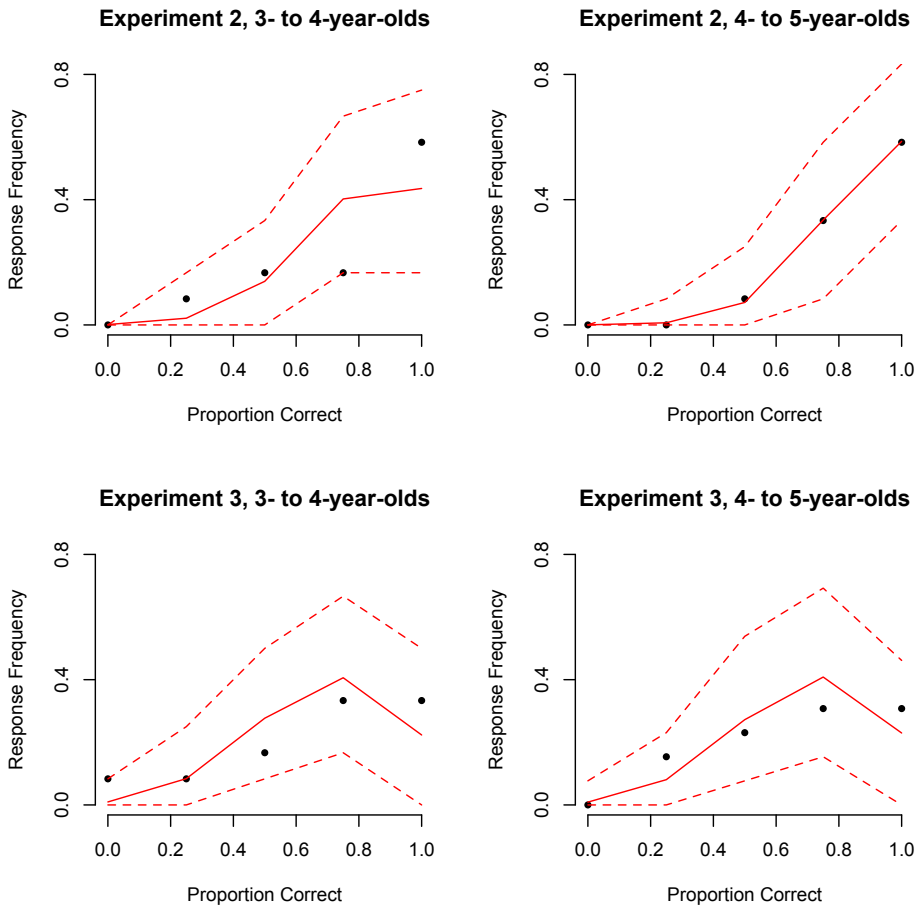
We thank the Editor again for the perceptive comment. In fact, there is evidence for coordination around a construct related to informativeness even in non-linguistic situations. In so-called "Schelling games," participants converge on a single solution (e.g. where to meet in New York City when a location has not been agreed upon) based on recursive reasoning about what the other person was likely to have chosen. So we agree that there is nothing *necessarily* linguistic about the kind of reasoning we are describing here. Nevertheless, we don't see evidence for the phenomenon we documented when we remove the linguistic cue (Experiment 4). We have modified the discussion of attention in the General Discussion and added a mention of non-linguistic coordination.

Reviewer #1

> *As noted by the authors (p. 21): "we cannot differentiate between the case in which each participants' judgments are slightly affected by aspects of the contexts and the case in which some participants notice the informativeness of a feature and others do not. This is a general issue in translating computational-level models of human cognition to the psychological process level (Frank, 2013)." While this may not be important (yet) in computational-level modeling, it is critical in understanding how children develop. All typically-developing children excel at inducing new word meanings. If using the informativeness of features is as important to learning new word meanings as the authors have argued, one would expect "each participants' judgments are slightly affected by*

aspects of the contexts" rather than "some participants notice the informativeness of a feature and others do not." Even though the proposed computational model does not yet make this distinction, the empirical data in Experiments 2 and 3 should be analyzed for individual children's response patterns across the 4 inference trials.

We thank the reviewer for this comment. In response to it, we have conducted a reanalysis of children's responses in the inference trials of Experiments 2 and 3. In each experiment, we compared the distribution of children's responses to the theoretical expectation if they were responding according to a binomial distribution (that is, choosing the correct answer with probability proportional to the population mean). The results of this analysis are shown in the figure below. The black points show the frequency of a particular response profile (proportion of inference trials correct out of the total of 4), while the red solid lines show the theoretical profile given a binomial distribution and the population mean. The red dashed lines show 95% confidence intervals on this theoretical profile. As you can see from this figure, in no case can we reject the hypothesis that participants answered according to a simple binomial choice. Had some participants used informativeness while others did not, we should have observed a bimodal frequency distribution that would have placed some frequencies outside the confidence bounds on a standard, unimodal binomial distribution. We haven't mentioned this analysis in the text, but are open to including it if the Reviewer or Editor would like.



Moreover, if using the informativeness of features is as important to learning new word meanings as the authors have argued, the tendency to use such informativeness should correlate with the children's vocabulary size. It will be helpful to replicate Experiment 3 and add a vocabulary measure. While this may not be relevant to testing the proposed computational model, it directly speaks to the real-world significance of the proposed model. That is, does young children's use of informativeness of features really make a difference to their learning of new word meanings?

We appreciate this suggestion from the Reviewer and agree that an important measure of any new factor involved in word learning is its contribution to overall vocabulary growth. But there are two obstacles that would prevent us from carrying out the proposed experiment, one theoretical and one practical.

First, on the theoretical side, we do not believe that every word—or even a majority of words—are necessarily learned through pragmatic reasoning. Good candidates might include some kinds of property terms, and perhaps super- and subordinate-level category terms in some cases. These terms make up a minority of children's early vocabulary and may well be taught explicitly (e.g. "a poodle is a kind of dog," as discussed in e.g. E. V. Clark's 2002 book). So, unlike speed of processing (Fernald et al., 1998) or even the shape bias (Smith et al., 2002), we might not predict large differences in vocabulary composition based on individual differences in pragmatic ability. But this is not the same thing as saying that pragmatic inferences are unimportant: there may be critical cases where words are learned faster or more easily due to the kinds of inferences we described in our manuscript, even if the eventual learning is overdetermined.

Second, on the practical side, the classic design for a comparison between some factor in word learning and overall vocabulary uses an individually-reliable measure of word learning and a parent-report measure of vocabulary such as the MacArthur-Bates Communicative Development Inventory (MCDI). Unfortunately, neither of these tools are available to us. With only four target trials per participant, our informativeness measure is not individually reliable—and there is some reason to believe that we might not want to add more trials because of the risks of learning during the paradigm. In addition, the age range of the children we tested is beyond the diagnostic range of the MCDI. While there are of course vocabulary measures that can be used with preschool children, they are longer and more difficult to administer and may not have the same relationship with particular aspects of word learning (depending on the makeup of the vocabulary being tested).

To summarize: Our argument was that preschool children can use their burgeoning pragmatic abilities to make inferences based on a presumption of speaker informativeness. We do not yet know the extent to which these inferences are important in the growth of vocabulary more generally (and have noted this in the revision). This question is unarguably important but for both methodological and theoretical reasons we believe it is a question for future work.

To the children in a university lab preschool, the experimental conditions were likely to be taken as guessing games with right and wrong answers. But in everyday life (e.g., toy stores, preschool classrooms), if there are several similar toys with different accessories,

an adult may call children's attention to the most interesting/outrageous type of accessory, even if there is more than one token in the array. So, if Experiment 1 had been done using photos of such everyday settings as just described, instead of the very orderly schematic drawings, would the results have turned out to be like those reported in this manuscript? What are Experiments 2 and 3?

…

This concern highlights a more general issue. Namely, what are the situations where the model is likely to predict well, and what are the situations where it may break down? A more in-depth discussion is needed to make this model more useful for understanding children's amazing word-learning feats in everyday life.

We thank the Reviewer for this comment. Based on this comment and those of the Editor, we have added discussion of the strengths and weaknesses of our model to the General Discussion as well as some links to new work that begins to address these issues (Smith, Goodman, & Frank, 2013; Vogel et al., 2014).

Figure 3: line graphs in developmental psychology typically are used to track within-child changes across ages (or across conditions). Bar graphs will be more appropriate for between-group comparisons here (age 3 versus age 4).

This figure has been replaced with a bar graph in the revised manuscript.

Reviewer #2

p7 "Nevertheless, accounts differ considerably on the age at which children first succeed in making implicatures (Papafragou & Musolino, 2003; Guasti et al., 2005) and on the factors that prevent them from succeeding (Barner & Bachrach, 2010; Barner, Brooks, & Bale, 2011; Stiller, Goodman, & Frank, under review)." -- Kurumada (2013) provides further evidence that --under the right task conditions-- kids much younger than in previous studies can and do draw pragmatic inferences. In addition, Katsos & Bishop (2011) show that 5-to-6-year-olds are sensitive to informativeness violations in Quantity implicatures.

We thank the Reviewer for these references, now cited in a slightly-expanded paragraph on pragmatic inferences.

p. 8 "Related predictions can be derived in a game-theoretic framework for pragmatics (Jaeger, 2010)". Also cite Franke (2009), "Signal to Act", whose work provides one of the most detailed spelled-out game-theoretic frameworks for pragmatics.

Thank you for mentioning this work, which was an inspiration for our original model – we have included a larger number of citations to game-theoretic pragmatics in the revision.

p13 It seems that two out of the four data points have confidence intervals that do not include the model's predictions. That should be discussed. (see also p. 15: the large R^2 is hardly surprising for four data points as long as the order correctly. To derive

predictions about the order of the four conditions, it would, however, not have been necessary to introduce a formal model. I thus thought the authors were aiming for a stronger point, namely that the model they propose provides a good quantitative fit. For that point, it's necessary to discuss why two out of the four data points do not quite seem to go with the specific predictions.

We thank the reviewer for bringing up this point, which is a general issue in fitting computational models to data: whether they should produce a fit that is high in relative or absolute terms. A high relative fit is captured by a correlation with human data: correlation measures the linear relationship of datapoints, irrespective of scale. A high absolute fit is captured by a measure of absolute distance (e.g. root mean squared error). While high absolute fits are of course preferable, they typically require adjustment of the data to match the vagaries of human performance (e.g. Frank, Goldwater, et al., 2010): When a model predicts 100% performance, we should assume that human learners will reach a high level of performance but perhaps not 100%. Adjusting models for humans' imperfect performance is common in many kinds of models and critical to achieving good fits even in domains like psychophysics (Wichmann & Hill, 2001).

As much as can be determined by four data points, the quantitative correspondence between our model and the data in Experiment 1 is quite tight (within the confidence intervals for every point), and the relationship is linear, but the slope of the line is slightly less than one (indicating the kind of imperfect performance described above). In the previous figure we had included a reference line with slope 1, which gave the impression that there was a larger mismatch than we believe there is; a modified figure with a simple linear fit is given in the revision. We also discuss this issue briefly in the results section for Experiment 1.

p13 "no random effect structure was warranted because each participant contributed only a handful of trials" -- that's a somewhat odd way of putting it. It's not that random effects are warranted. It's just that under the required model, it would be hard to a) reliably estimate the random slopes for participant and b) to disentangle the effect of interest from random differences between participants (i.e., the effects would likely lose significance). That's a shortcoming of the design and should be acknowledged as such. Jaeger et al (2011) discuss problems that can arise when trying to fit random slopes to data with few observations per grouping factor level.

We agree, and have modified this statement; the revised discussion is in a footnote and references Jaeger et al. (2011).

p17 "we fit a logistic mixed effects model to children's responses, with age group and condition as fixed effects, and with random effects of condition fit for each participant and each target item (Barr, Levy, Scheepers, & Tily, 2013)." -- if the authors meant to say that their model contained the maximal random effect structure, that should be stated more clearly. Additionally, maximal random effect structures for logistic models can be problematic if there are too many 0 or 1 cells in the subject x item x design table. See also Jaeger et al. (2011) for a discussion of problems that can arise when fitting random slopes to data with few observations per grouping factor level (see my comment above).

Finally, it seems Jaeger (2008) introduced the method to cognitive psychology (and uses maximal random effects). Perhaps cite that work instead/as well.

We did mean to state that our model included maximal random effect structure; we have revised this sentence to make this statement and provide clearer citations.

same page "A model with an interaction term did not provide better fit (c2(1) = :16, p = :69)." -- inclusion of the interaction term could still change the significance of the main effects. It should be stated whether that was the case.

The interaction term reduced the significance of the coefficient on filler trials but did not alter the pattern of results for inference trials. This is now noted in the manuscript.

p. 18 "This exclusive purpose may have given participants a greater sense that the utterance should be chosen with maximal informativeness". It is not clear why the exclusive purpose should have had that effect. Elaboration would be helpful. In general, the motivation for Experiment 3 is somewhat unconvincing/unclear.

We have added some elaboration regarding why an exclusive purpose would lead to a greater presumption of informativeness.

p. 19 How do the authors explain the interaction of trial type and age group? That is, what is the explanation for 4-year-olds' superior performance on filler trials?

Our explanation is simply that 4-year-olds often have a tendency to follow directions better and tend to perform better on many kinds of tasks (there is a trend towards this pattern in Experiment 2 as well).

Finally, depending on which direction the authors plan to take for this paper, they could consider covering the adult literature on pragmatic inference more (perspective-taking in adults, as in work by Keyser; Barr; Heller; scalar inference/implicature in work by Snedeker; Grodner; Degen; and others; the former is briefly mentioned in the discussion, but the authors could go further: are there existing quantitative results that the model would correctly predict?).

We thank the Reviewer for these references. We are aware of this literature and find it very interesting, but believe that it is outside of the scope of the current manuscript, which focuses primarily on acquisition. In the version of our pragmatic model that focuses on reference resolution rather than word learning, these are the most important current findings to fit. But we think that discussing the adult scalar implicature literature in more depth here would diffuse the message of this paper. We're happy to revisit this decision if the Reviewer or Editor think otherwise.

Reviewer #3

1) A concern: An alternative explanation of the findings is that they tell us nothing about

language and instead only tell us about how participants view objects contrastively. Consider E1, where participants are asked to make bets about the meaning of an adjective and the bets pattern with informativeness. To say that the results tell us about how adults infer the meanings of adjectives per se, we would first need to show that the same pattern of results does not obtain if the language aspect of the task is absent. Consider if we changed the task in E1, dropping the adjective all together, and simply asked the subject to value each of the two features of the alien based on "how well they like it". If the findings still patterned the same, with higher bets for more informative features (I bet they would), would we still want to conclude these findings tell us about language? I think not. In that case, I think it would tell us more about perception of contrast sets. Similarly, if the child experiments dropped the adjective completely (e.g., "look at this one" in training, and "find another one" at test), would we still find the same results? Without doing this test, we can't be sure the observed findings are about language per se, and not due to something else about task.

We took this suggestion very seriously and ran Experiment 4, which includes a "salience" control of the type described by the reviewer. We find that the effect is not present when no word is used. (Note that we used a common "show me another" framing, rather than the preference framing suggested above). In addition, we find that children are in fact making use of the *specific* label that is presented in training, as changing this label at test reveals a very different pattern of performance (Experiment 4, "disambiguation" control). All in all, we believe that these new data provide strong evidence that the effect we describe is truly due to the use of a particular label by the experimenter, and not to more general familiarity or salience of the unique object.

2) A suggestion for improvement: In Experiment 1, the task would seem to simplify the real world problem by highlighting minimal pairs with unique members identifiable by a restricted class of attributes. The task thus teaches the learner what possible word meanings are, constraining the relevant space of possibilities away from potential meanings such as "poetic" or "vacuous". More concretely, for example, the alien pictures are all "green", they're all "eyed" they all have "tails", and they all have feet and noses, so in some sense, the conditions aren't really 1/1, 2/3, 1/2 and 1/3, but instead closer to 1/6, 2/8, 1/7, and 1/8, and if we include things like "vacuous" in the mix, the denominator is clearly infinite. But the cool thing here is the math shows us that the participants are doing the calculations based on the task-relevant features. So not only are they figuring out informativeness of adjectives and placing bets on that--they're defining informativeness within a restricted range of possible features as defined by the task (e.g., which of all possible features are interchangeable and referencable attributes). This aspect of the finding is cool, and I think deserves note.

Interesting point – we agree with this point in principle. But note that the dependent variable in our experiment is fundamentally a choice between two options. So even if vacuous predicates could in principle bias the computation in a free-response task, the restriction of alternatives would restrict our computations in the 2-alternative betting task.

p3: Gleitman & Gleitman (1992) have a catchy summary of the problem ("a picture is worth a thousand words and that's the problem").

p4: Brown-Schmidt & Tanenhaus (2008) have a nice example of how task-related goals eliminate ambiguity in unscripted conversation. They also show that the rate of underspecified NPs is high and necessarily contextually interpreted.
p7: There's actually a good deal of evidence for sensitivity to partner knowledge in young children. Citations include: Matthews, et al. (2006; 2007; 2010); Nilsen, et al. (2008); Nilsen & Graham (2009); Scott, He, Baillargeon, & Cummins (2012)--also see references therein & discussion regarding explicit vs. implicit response measures.

Thank you for these suggestions. We have incorporated many of them into the revision.

p12: I'm having a little trouble imagining what exactly the other conditions look like so it would be helpful to have a figure like the left half of Figure 2 that showed the four conditions with what an example target+context would look like in each.

Added, thank you for the suggestion.

For the purposes of the task, how were the two features of the target object referenced to the subject for betting purposes? Was there a separate disembodied picture of the accessory they put a value next to? Or were the features labeled somehow? If so, how?

There was a separate picture of the accessory next to the text box; this detail is now noted.

p15: It would help to have pictures of the stim for the two conditions.

Added, see above.