Justine T. Kao,
Ph.D. Candidate

Stanford University
Department of Psychology
450 Serra Mall, Jordan Hall
Stanford, CA 94305-2130

Dear Editorial Board,

We thank the editor and reviewers for their very helpful comments. We have attempted to address each of the concerns raised, either clarifying or revising as appropriate. We believe that the manuscript has improved significantly as a result. Below we respond to each point, indicating changes made to the manuscript where applicable.

If you have any questions or require any additional clarifications, please do not hesitate to contact us.

Sincerely,

Justine T. Kao

Editor's comments:

(1) Clarify the situation with respect to the case of 50 vs 51, or 100 vs 101, in Fig 1, where the exceedingly small experimental difference that is reported here seemed very surprising to both reviewers, and to me, and would seem prima facie to disconfirm the model to some extent.

> In fact, Figure 1 represented model predictions rather than experimental results. We have revised the caption and text to more clearly reflect this. Thus the small differences between "round" and "sharp" numbers in Figure 1 is not evidence (for or) against the model. One may then wonder why the model predicts small differences and whether this is confirmed by the experimental results.
>
> The size of the halo effect predicted by the model is influenced by a cost parameter (see Eq. 7), which determines how much more "difficult" it is for a speaker to utter "51" versus "50." This difficulty may be due to factors such as availability, frequency, and complexity of the number terms, as suggested by previous research (Oldfield & Wingfield 1965; Balota & Chumbley 1984; Mehler 1992; Jansen & Pollmann, 2001; Solt et al., 2011). However, we do not make assumptions about which particular psychological factors are responsible for this cost. We simply assume that there is a cost difference, and show that a model incorporating cost can accommodate human data. That is, we treat cost as a free parameter, which we fit to the experimental data.
>
> We originally posited a sharp/round cost ratio of 1.8 with the guidance of intuition and model experimentation. Based on the helpful comments from the editor and reviewers, we decided to fit the cost parameter in a more systematic manner, as now described on page 6:
>
> > "We jointly fit λ and the model's cost ratio C to optimize correlation with the behavioral data. The best fit was with λ =0.36 and C=1.3, resulting in a correlation of r = 0.974 (95% CI = [0.9675, 0.9793]). The range of cost ratios that

produces correlations within this confidence interval is [1.1, 3.7], which is quite broad, suggesting that the overall model fit is not very sensitive to the cost ratio. To further capture the details of the halo effect, we jointly fit λ and C within this range to a measure that is more sensitive to utterance cost: We computed the difference between the probabilities of exact versus fuzzy interpretations for each utterance, which gives us each utterance's bias towards exact interpretation. We then computed the difference in this bias for sharp versus round numbers, which gives us a "halo" score for each sharp/round pair. We fit λ and C to minimize the mean squared error between the model and humans' halo scores. We found that the cost ratio that best captures the magnitude and pattern of the halo effect found in participants' data is 3.4, while $\lambda = 0.25$. This produces an overall correlation of 0.9677 with human data from Experiment 1."

This two-step fitting process finds the parameters that best match the overall pattern of the data, but then best matches the extent of the halo effect, within the statistical uncertainty of the broader measure. Since the optimized cost is now higher, this more systematic selection of the cost parameter yields a larger difference between the model's exact/fuzzy interpretation of "50" and "51," as shown in the revised Figure 1.

Figure 3b summarizes and quantifies the extent of the halo effect for both experimental data and model predictions. We see that the halo effect, where sharp numbers are generally interpreted more exactly than their round counterparts, is significant, though small, in the data and that the model captures both the pattern and overall magnitude of the effect. Hence, the model is able to accommodate the halo effect in our data via the cost parameter, which we interpret as aggregating many factors that could differ across situations. We expect that the halo effect, and hence the contextually appropriate cost parameter, could be larger in other situations for reasons including those that Reviewer #2 suggests (such as using Arabic notation, certain characteristics of the dependent measures, etc.).

(2) Make the clarifying introductory remarks about the model that Referee #2 asks for.

We have revised the introduction to include an informal overview of the modeling approach and our novel contributions (page 1, second paragraph). We have also more clearly described the RSA model and its components where they are introduced and connected this to the later discussion of the model.

(3) Make the charts clearer, more self-explanatory, and easier to read in the way that Referee #1 asks for.

We have made the figures more self-explanatory and easier to read by including more detail in the captions and increasing the size of the labels. We have also made the axes labels more self-explanatory and easier to interpret. Finally, we have changed the colors and shapes so that the figures are readable when printed in grey scale.

(4) Say more about the prior probabilities, where they come from (just stipulated? how might one do better?),

> A main contribution of the modeling framework used here is that it formalizes the relationship between non-linguistic background knowledge and the structure of language in giving rise to interpretations. Background knowledge is captured via prior probabilities, which the model depends on (but doesn't predict). We measured the necessary prior probabilities of prices and affect empirically in Experiment 3a and 3b, respectively. The main text (p. 2) has been revised to emphasize that we measured the priors by asking participants to report the probability of prices (Experiment 3a) and affect given prices (Experiment 3b). We have also clarified that we believe separately measuring non-linguistic background knowledge and linguistic interpretations (as we have done) is the best way to isolate the structure and dynamics of language per se, for example in the following:
>
> > "We address the third principle [of communication] by empirically measuring people's background knowledge to understand the interaction between nonlinguistic and linguistic knowledge in shaping language understanding." (p.1)
>
> We could have established the prior probabilities in a different way by examining a database of prices from sources such as Amazon.com, in which case we would be able to measure "true" price distributions instead of people's estimations of these distributions. However, eliciting people's beliefs directly allows us to more immediately assess how people use their (potentially noisy and imperfect) world knowledge to understand language. (Indeed, a preliminary analysis of prices scraped from the internet suggests that people have rare but systematic deviations from natural statistics in their beliefs about price frequency – for instance people under-predict the prevalence of expensive watches.)

How important a role do the priors play in the predictions -- each reviewer had some questions about those.

> The prior probabilities play a very important role in the predictions. As shown in Figure 3a and Figure 4b, replacing these priors with a uniform distribution alters the model's predictions, which then fail to produce the patterns we see in participants' interpretations. By eliciting people's background knowledge of these distributions and incorporating them in our model, our goal was to show that both nonlinguistic knowledge (such as the probability of certain states of the world) and linguistic knowledge (such as the literal semantics of "1000" and general principles of communication) shape how people understand language. Our model is able to combine these types of knowledge to accurately predict people's interpretations of number words.
>
> The prior probabilities are critical to the model's success. However, the prior probabilities alone are not sufficient, since relying on prior probabilities alone without taking into account the structure of language would result in identical interpretations of "The watch cost 50 dollars" and "The watch cost 1000 dollars." By incorporating both

types of knowledge, the model interprets the utterances in ways that closely align with human judgments.

And some optional but recommended changes:

(5) Try to address the "significance-skepticism" expressed by referee #1 in a way that could help non-linguist readers better appreciate the value of these results and the novelty of the methods used.

> We believe that the skepticism expressed by the reviewer plays out at two levels: narrowly, the extent to which our work represents an important extension to previous formal models and informal ideas, and broadly, the value of formalizing and quantifying linguistic intuitions. To address the former we have revised the main text to emphasize that our contribution consists of more than an obvious extension to the previous RSA models (see page 1 second paragraph and page 2 first full paragraph). We provide more detailed discussion of these points below. Engaging deeply with the latter, broad, skepticism may be more than we can accomplish in this article, though we have adjusted the wording of the conclusion to emphasize the explanatory value we see in formal and quantitative work on language understanding.

(6) Address any of the other concerns of the two referees that you are able to.

> We have done so: please see comments to specific reviewer items below.

Reviewer Comments:

Reviewer #1:

Comments:

In many ways I think this is a nice paper. The modeling is solid; the findings are well-explained; and the core result -- that nonliteral interpretations like pragmatic halo and hyperbole follow naturally if you assume that both speakers and listeners are making interpretations about the affective subtext of an utterance -- makes complete sense. The paper should definitely be published somewhere. The core result, while it makes sense, would also come as a surprise to very few people, especially people who study pragmatics. In fact, the core result is already suggested by many existing theories of nonliteral language use, and supported by lots of other empirical data.

> We believe our core result is more than the (pre-existing) observation that hyperbole conveys affect or that conversational partners are reasoning about affect: it is a detailed formal analysis of how this reasoning proceeds, and how various factors interact to give rise to interpretations, which is supported by quantitative agreement with experimental evidence.

We believe that a detailed formal analysis is crucial because there are alternative analyses that fit under the general theme of reasoning about affect. Indeed the more obvious analysis, which we started with and compare to, simply extends the meaning representation to include an affective dimension. We incorporated a two-dimensional representation of meaning in the model to capture both the state of the world and a speaker's affective attitude towards it. As Reviewer #2 noted, this is related to Chris Potts' work regarding the expressive dimension of language (Potts, 2007). However, simply extending the representation of meaning to two dimensions without including communicative goals (aka conversational topic, aka question under discussion) is insufficient for modeling the nonliteral interpretations that we see in our behavioral data. In a model where the listener considers both the state of the world and affect but does not reason about which dimension the speaker wants to communicate, the listener would infer that "The kettle cost 10,000 dollars" means that the speaker likely thinks it was too expensive, because a $10,000 kettle is a priori associated with a high probability of affect. Such an extended model would be able to capture information about the speaker's affect. However, the listener under this model would still interpret the utterance to mean that the kettle actually cost $10,000, because the listener believes that the speaker wants to be informative, and there is nothing to be informative about except the actual price state. To fully explain and model nonliteral interpretation, we incorporated a second critical insight, which is that the listener needs to reason about the speaker's communicative goal, namely which dimensions—price, affect, or both—the speaker wants to communicate about. By jointly inferring the speaker's communicative goal and the price state and affect, the listener can now reason that "The kettle cost 10,000 dollars" is a very likely utterance given that the actual price is around $50, the speaker thinks it's too expensive, and the speaker only cares about maximizing information regarding affect. This results in an interpretation that is much closer to people's judgments.

Our insight regarding communicative goals is closely related to previous theoretical and empirical work showing that context and questions under discussion shape people's interpretations of sentences (e.g. Wilson & Carston, 2006; Duranti et al., 1992; Garrod & Sanford, 1994). However, to our knowledge our work is the first to formalize this insight and the first to provide concrete evidence that both elements are crucial for producing nonliteral interpretations. It is not clear from previous work, at least to our knowledge, that the listener's uncertainty about the question under discussion is a critical part of what drives nonliteral interpretation.

It is also far from clear from previous theories how various factors – background knowledge, literal meaning, affective dimensions, question under discussion – should be integrated together in language understanding to give rise to the particular, graded interpretations that people arrive at. Indeed, we believe that quantitative theories and data go hand-in-hand in shouldering an explanatory burden not addressed by informal theories: they achieve a level of precision which helps us move beyond the illusion of understanding engendered by 'obvious' intuitions or the, real but coarser, understanding enabled by informal theories and qualitative data.

[need to re-read this section later – may be kind of ranty.]

Thus, the contribution of the paper is not to be the first to suggest that nonliteral language use conveys emotion and emphasis (i.e., affect): many people already believe this. The contribution of the paper is simply to formalize this idea. The formalization, moreover, is a fairly straightforward extension of an existing series of models (the RSA/pedagogical models). The extension consists of simply adding the existence of an affect (and a corresponding prior P_A) into the recursive structure of the model. It then derives people's utterances and interpretations by assuming that both people involved in the conversation reason about the affect as well.

> As described above, simply adding an affect dimension to the recursive structure of the model is insufficient for capturing nonliteral language understanding. As shown in Figure 2a, a model that has the affect dimension but does not reason about the speaker's goals interprets "The electric kettle cost 1000 dollars" as meaning that the kettle actually cost $1000 dollars, which is clearly not how people interpret the utterance. This is because while the "No goals" model is aware of the affective dimension, it is not able to reason about which dimension is more likely to be relevant, and thus is not able to reason about which dimension is unlikely to be literally true. Instead, the structure of recursive reasoning must be adjusted such that the pragmatic listener is uncertain about the topic of conversation, but believes the speaker knows it. Incorporating goal (or QUD) inference in this way is a nontrivial and important extension to RSA, and may be critical for predicting a range of rich phenomena in language understanding.

A model should, ideally, tell us something new, or at least clarify and extend existing beliefs.

> We have attempted to revise the main text to highlight the novel contributions, clarifications, and extensions contributed by the model presented in this paper (see introduction on page 1 and discussion on page 5).

On less subjective matters:

1. I really didn't see how the model captures pragmatic halo. (That is to say, it didn't look to me like it did, but I might be missing something). According to the main text, Figure 1 apparently shows the pragmatic halo effect, but as far as I can tell, it doesn't - at most it shows an extremely small effect that is hardly visible on the graph at all. Looking at the "fuzzy" row (which I believe corresponds to the times that the model interpreted the utterance as "fuzzy"), it appears to me that for all the price pairs (e.g., 50/51, 100/101, etc) there is no difference between the probability assigned to the exact one and the probability assigned to the fuzzy one. Why would an exact utterance like 51 be equally likely to be interpreted by the model as fuzzy as an utterance like 50? Similarly, in the "exact" row the probability that the model interpreted it as "exact" if it was 51 is only very very slightly more than if it was 50 (this difference is hardly visible at all). This certainly doesn't seem to capture the effect very strongly.

> Please see our response to the editor's comment (1). We believe it is now clearer that (and how) the halo effect is captured by the model.

[Relatedly, are the axes for the model in Figure 3b the same as for humans? If so I'm quite surprised given Figure 1, and I think I must be misinterpreting Figure 1. If not, then it really needs to show the model axes as well because otherwise it may be quite misleading about the magnitude of the predictions made by the model vis-a-vis the magnitude of the effect in humans.]

> The axes in Figure 3b are identical for the model and for humans. The magnitude of the predictions made by the model very closely aligns with the magnitude of the effect in human judgments. What may have been confusing in the presentation is that the axes in Figure 3b is the *difference* in probability between an exact interpretation and a fuzzy interpretation of an utterance, while the axis for Figure 1 is the absolute probability of exact and fuzzy interpretations. We have changed the axis label for Figure 3b to make this clearer. Also, note that the scales for Figure 3b are relatively small and fine-grained (0.04 ~ 0.10, 0.02 intervals), while the scales for Figure 1 are larger and more course-grained (0 ~ 0.8, 0.2 intervals). We have adjusted the scales for Figure 1, so that the magnitudes in Figure 1 and Figure 3b are visually more comparable.

2. A key question I have reading this paper is how much of the model performance is due to the prior probabilities P_A and P_S.

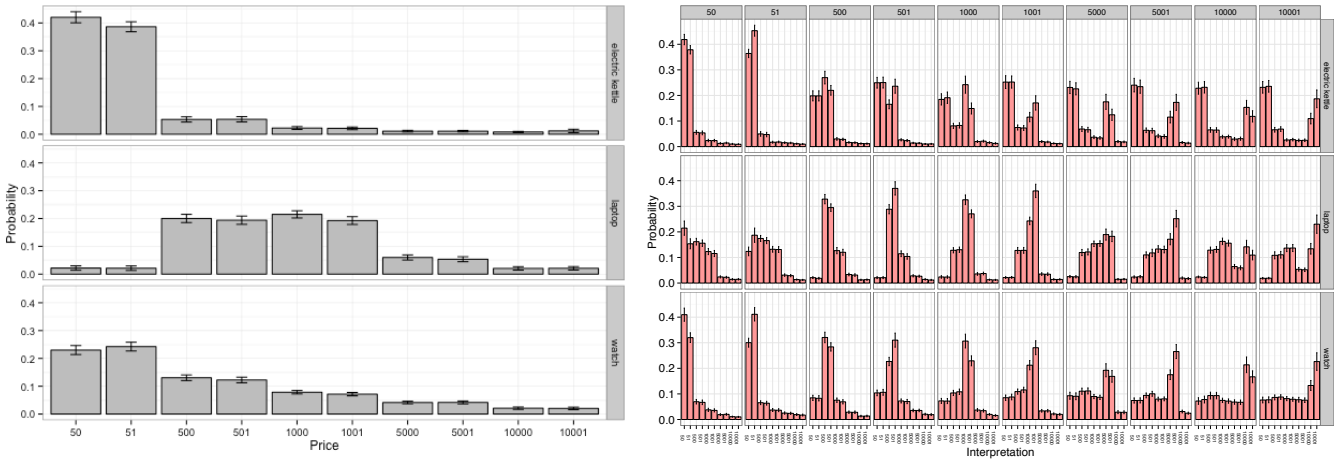
> As described in our response to the editor's comment (4), the prior probabilities P_A and P_S are indeed extremely important for the model's performance. This is appropriate because we believe the background knowledge captured by these prior probabilities is extremely important for the language understanding process.

Put another way, how much of this could be explained assuming a model that didn't have this recursion (i.e., just stopped after modeling the speaker assuming a literal listener, i.e., stopped at the base case of the recursion)?

> Our model describes the listener's interpretation distribution and not the speaker's utterance distribution. The base case (with no recursion) then is the literal listener. The literal listener modeled by RSA incorporates background knowledge to weight interpretations that are possible given the literal meaning; but this listener has no means by which to arrive at a meaning inconsistent with the literal meaning. Indeed, it is the addition of the goals (or QUD) and the inference about these goals, which can first happen at the level of the pragmatic listener, that enables non-literal interpretation. If we stop at the literal listener, even including the same prior knowledge, "10,000 dollars" would always be interpreted as meaning $10,000 with probability 1. As a result, the recursive step in which the listener reasons about a speaker and, importantly, the speaker's goals, is necessary for producing hyperbole.

More generally, it would also be interesting to just see the priors for each, because I have the strong impression that a great deal depends on having the correct ones (indeed, Figure 4 suggests this).

A great deal does depend on having the correct priors, as explained in the response to the editor's comment (4). However, the priors alone are insufficient. Below is a figure showing the price priors. Comparing this figure to people's interpretations of the different utterances (Figure 5b in the paper), we see that while the priors affect people's interpretations of utterances, participants assign different probabilities to states that are extremely unlikely under the prior distribution given different utterances (e.g. participants are much more likely to interpret "10,000 dollars" as meaning $10,000 than to interpret "50 dollars" as meaning $10,000). This means that both nonlinguistic information from the price priors and linguistic information from the utterance shape people's interpretations. However, it is not obvious how one would combine these two sources of information in a straightforward or linear manner. Our model provides a natural and theoretically motivated way of incorporating both sources of information.



The way P_A is constructed, in fact, as it is almost builds in the answer: a set of participants has rated the prior probability that an object costs $X, as well as the probability that people think it's expensive at $Y if they say U when the talk about it... and then, a model that knows this concludes something sensible about the price of the object that the person says U about.

To clarify, in eliciting P_A participants judged whether an object costing a given amount ($Y) is expensive *in the absence of any utterance U*. The main experiment then explored the interaction of linguistic evidence (in the form of an utterance) with this non-linguistic background knowledge.

That is (sort of) fine from an explanatory viewpoint, because presumably adult speakers do have these priors and that is what they are using to make sense of these interactions. But in another sense it does leave a huge amount unknown: how do people learn the priors - especially P_A? (It's fairly obvious where people might learn about P_S, i.e., prices of kettles, laptops, and watches).

How people learn priors is indeed an interesting and important question that deserves further research, though it is beyond the scope of the current paper. Instead, we assume

that people have certain knowledge of the world, and our goal is to measure that knowledge and incorporate it in a model of pragmatics to show that it predicts how people interpret utterances. We agree with Reviewer #1 that P_S could be learned by exposure to the prices of various everyday items. As for P_A, we speculate that after having learned the price distributions of various items, people develop a judgment for which prices are unusually high and are likely to elicit an affective response, given that people generally do not want to pay an unusually high amount of money for an item. While there may be a way to construct P_A from P_S for the different item types, we decided that it was much more straightforward and accurate to directly elicit P_A from subjects.

How much of the theoretical work is it doing? Given that I suspect the answer is "a lot" it would be at least nice to see what it is, and how similar it is to the model predictions.

Please see above and our response to editor's comment (4).

3. In general these figures are very hard to read. I appreciate the authors are trying to put a lot of information in a little space, but the captions could be way more informative. For instance, the meaning of the axes is never explained for Figures 4a and 2a - going into the text I see it is the likelihood that the item cost that much (but how much? The same as the utterance? I am baffled. And I shouldn't have to scour the text just to be able to read the figure).

The captions for Figure 2a and 4a have been revised to clarify the meaning of the axes. Both the x and y axes are in units of probability. In Figure 2a, each point in the scatterplot represents the probability that a certain utterance is interpreted as a certain price state. Since there are 10 utterances for each of the 3 item types, there are 30 possible unique utterances such as "The electric kettle cost 50 dollars" and "The laptop cost 10,000 dollars." For each of the utterances, there are 10 possible interpreted pries ($50, $51, $500, $501, etc). As a result, there are 300 possible utterance/price state pairs, and thus 300 points. Figure 2a plots the probability of each utterance/price state pair as predicted by the model (x axis) and rated by humans (y axis). Thus the position of a point (which is an utterance U/price state S pair) is determined by the probability that the utterance U is interpreted as price state S.

In Figure 2b, each point represents the probability that a certain utterance is interpreted as conveying affect when the interpreted price state is greater than or equal to the utterance (in other words, when the utterance is either hyperbolic or literal). There are 45 such utterance/price state pairs (collapsed over round and sharp numbers), and thus 45 points in Figure 4a. Figure 4a plots the probability of each utterance/price state pair conveying affect as predicted by the model (x axis) and rated by humans (y axis). Thus the position of a point (which is an utterance U/price state S pair) is determined by the probability that an utterance U interpreted as price state S conveys affect.

It would also be nice to make figures that show up if the paper is printed in grayscale - e.g., use different shapes and brightness levels for the points in the figure, not things that are so similar

that they all look the same in gray. Also, the labels on most of the figures are tiny and everything is quite hard to see.

> We have adjusted the figures accordingly (see our response to the editor's comment (3)).

4. I know this is in the realm of further model development... but why do we need to have non-uniform utterance cost to get the bias for exact interpretation in Figure 3b? I mean, I see how the model requires it - but that seems like a kludge to me: why should "51" be more costly to say than "50"? If it's simply length of time it takes to say it, then, "250" should be way more costly than even "51" but I would suspect you don't want to assume that. (I know there is a literature with other models that assume this but I think it's a kludge there too).

> As described in our response to the editor's comment (1), the cost of an utterance may be due to factors such as availability, frequency, and complexity of the number terms, and we do not make specific assumptions about which factors are most important. We have revised the main text to make it clearer that cost is not necessarily determined by the length of the utterance.
>
> As the reviewer points out, similar assumptions about differential cost are widespread; we are open to the idea that they are all 'a kludge'… and would be interested to see an alternative! Is this too aggressive..?

You shouldn't have to hardcode a cost into a full model: it should follow from a properly specified affect (perhaps you would need one with more than binary states). That is, the model should reflect the intuition that people will say "50" instead of the actual price of "51" not because 51 is costly for some reason, but rather to (a) communicate uncertainty or lack of confidence about the exact cost - I know there was a CogSci paper in 2011(?) about this but I forget the exact title, sorry, maybe something about number preference - or (b) if more precise information isn't relevant. A model that included these factors should get the pragmatic halo effect to follow naturally, rather than getting an effect [to the extent there is one - see my #1] for a more uninteresting reason.

> While there may be a way to use a more fine-grained representation of affect to capture the different degrees of affect conveyed by sharp versus round numbers (e.g. confidence levels, etc), our model posits that pragmatic halo is primarily driven by utterance costs, while hyperbole is primarily driven by affect. What unifies the ways in which the two phenomena are modeled is the fact that in both cases, the listener reasons about the speaker's communicative goals and performs joint inference on the goal and the meaning. Our model thus shows that incorporating goal inference allows us to flexibly integrate different types of linguistic and nonlinguistic knowledge in the listener's reasoning. We believe that this is a rather interesting and novel explanation for the pragmatic halo effect, and we hope that our revision reflects this more clearly.
>
> It is worth noting that the symmetry between round and sharp numbers (50 vs 51) must be initially broken in *some* way. Whether this is via a cost term or another contingent assumption about those lexical items may be a matter of taste.

Also, why was utterance cost set as it was, and how dependent were the final results on the particular value of C(u)?

> We have fit C more systematically in the revised manuscript. Please see our response to the editor's comment (1).

5. Finally, this is by no means a criticism, just a thought question for future work. It seems to me that this model (which accepts non-truthful interpretations by interpreting them as non-literal) would break if it ADDITIONALLY had the capacity to recognise that sometimes people lie: either it would always think that a non-truthful interpretation was a lie, or it would think that all non-truthful interpretations were hyperbole or pragmatic halo. (I might be wrong.) People, obviously, can make sense of both nonliteral interpretations and blatant lies. How would you try to reconcile both possibilities within a single model? (We can even allow the liar to not be very good - i.e., to be modelled by not taking the recursion very far). This could, I think, be very interesting (maybe even interesting enough for PNAS) because the answer is much more non-obvious (at least to me!).

> This is a really interesting idea and certainly merits further investigation, although perhaps not within the scope of this particular project. In our current model, we assume that the speaker and listener have full access to the same background knowledge P_S and P_A, i.e. common ground. In a model where lying is possible, it would be interesting to examine whether a listener that has only partial background knowledge would be able to identify when the speaker is lying. A crucial difference between interpreting an utterance as a lie versus a hyperbolic statement is whether the listener believes that the speaker means for the listener to uncover the true price state. When the background knowledge is fully accessible to both speaker and listener, it is easier for the listener to uncover the true price state. However, when the listener has uncertainty about the background knowledge, we do predict that interesting predictions about lies and hyperboles could emerge.

Reviewer #2:

Comments:
This paper achieves the following: It develops (a) a quantified theory of the approximate use of round numbers, as in we waited thirty minutes to get a table, and (b) a quantified theory of the hyperbolic use of numbers (e.g. we waited a million years to get a table). These two uses are examples for non-literal, figurative speech that have the advantage of being easily controlled. The paper gives a model for these interpretations within a model in formal pragmatics called Rational Speech Act models (derived from the iterated best response model of Jäger & Ebert 2009), it provides for the setting of parameters for this model by an experimental investigation using Amazon Turk, and it tests the predictions of this model in an experiment, again using Amazon Turk. This constitutes a novel application of the RSA model to two types of phenomena that are known in linguistic research and have only partially found formal explanations (in the case of the approximate interpretation of
round numbers).

We think that the content of the paper merits publication in PNAS. It provides a precise and novel account for familiar linguistic phenomena within a theoretical setting that has been widely used for other fields (signalling theory, evolutionary game theory, e.g. in biology), and hence should have an appeal to an audience beyond linguistics and communication theory. The scientific methods used are generally sound.

But we also think that the paper, as it stands, makes it difficult to follow, and so we encourage the authors to improve on its presentation. This concerns the explanation of the RSA model on p. 1, where crucial parts remain unexplained - e.g., C(u) for the cost of the utterance, the idea of recursivity, the role of e to create a "diminishing return" for each recursion which guarantees asymptotic behavior, P(m) for the prior probability for the meaning m, etc. These things are partly explained later, but at this point the reader is puzzled. Perhaps it would be suitable to give an informal overview and then integrate the presentation of the model with the application at hand, which is done here in the section "Model" in the "Materials and Models" section on p. 5.

> We have both expanded the informal overview of the modeling approach in the introduction and detailed the components of the RSA framework where they are first used (see our response to the editor's comment (2), above).
>
> While the RSA framework is defined to arbitrary recursive depth, we have followed previous research (Frank & Goodman, 2012; Goodman & Stuhlmuller, 2013) suggesting that human language understanding is often best fit at recursive depth 1. This makes it less relevant to explain the asymptotic behavior of the model. To clarify this we have adjusted our notation to the more specific case of one level of recursion that we use to model participants' judgments.

On p. 2 the authors suggest a model in which interpretation has two dimensions, one related to the state of the world, one related to the emotional attitude of the speaker. This is directly related to Christopher Pott's two-dimensional theory of meaning, which should be mentioned here.

> We thank Reviewer #2 for the reference, and we included it in the main text when introducing the two dimensions of interpretations. There are strong similarities but also critical differences between our mechanism and Potts's, hinging on the extent to which affect is affected by semantic operations, which could be interesting to explore in future work. (necessary?)

On Fig. 1, it is astonishing that the experimental results showed only very small differences between e.g. "100" and "101" in terms of likelihood for an exact interpretation. We expect an effect of the type of experiment here, see below.
What is the motivation to take 1.8 as cost for sharp numbers (with respect to 1 for round numbers)?

> Please see our response to the editor's comment (1).

In general, readers might be alerted that roundness does not always relate to shorter numbers, e.g. with the temporal scale (where e.g. 45 minutes can be argued to be rounder than 50 minutes,

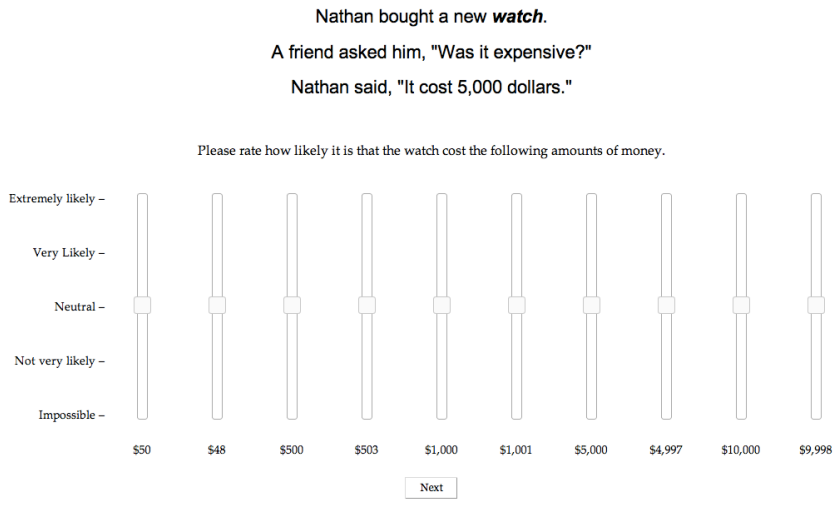
cf. Solt e.a., http://www.zas.gwz-berlin.de/fileadmin/mitarbeiter/solt/The_Preference_for_Approximation_-_final.pdf)

> As described in our response to the editor's comment (1), we believe that utterance cost may be due to factors such as availability, frequency, and complexity of the number terms. We agree that these factors are context-dependent, and so different numbers may be round or sharp in different contexts (such as prices, times, and number of eggs, etc.), as Solt et al. (2011) have shown.

There are two problems that we see with the experiment itself, not with its presentation:

1. It strikes us that the question how much a kettle, watch or laptop computer ("really") cost can be seen as questioning the truthfulness of Bob, the speaker. This is o.k. for the hyperbolic reading, but perhaps not so for the fuzzy reading.

> We believe the exact wording and presentation we used in Experiment 1 makes it less likely to sound like Bob's truthfulness should be questioned. The exact presentation is shown below. By asking participants to rate each of the possible prices on slider bars, and by wording the task as "Please rate how likely it is that the electric kettle cost the following amounts of money," we are not indicating that Nathan did not "really" spend $5,000 on the watch. That said, it is possible that the task context has an affect on interpretations, which is reflected in the relatively small value of the fit cost parameter.



Also, if a subject has heard Bob say a sharp price, like 51 dollars, then it is likely that he will be understood as specifying a sharp value when Bob used a round number, like 50 dollars.

> We thank Reviewer #2 for this insight and suggestion. One detail to clarify is that since the speaker's names are randomized and each name only appears once for each participant, it is unlikely that subjects will attribute biases towards using round versus

sharp numbers to a particular speaker. However, we agree with the general concern that there may be order effects in the data. To examine this, we took the first 5 trials that each participant saw (out of 15) and compared them with the last 5 trials. We performed a paired t test on the average probability ratings for each utterance/interpretation pair, broken down by the type of interpretation (exact, fuzzy, or hyperbolic). There were no significant differences in the probability ratings for any of the three interpretation types. There is also no significant interaction between the trial order (coded as first 5 trials v.s. last 5 trials) and interpretation types ($F(3, 592) = 0.4476$, $p = 0.7191$). This suggests that at least in this experiment, exposure to different types of utterances in the previous trials do not seem to have a significant effect on participants' interpretation of utterances in the later trials.

2. If numbers are given by the Arabic notation, as in 50 dollars, then it is likely that there is a bias towards a precise interpretation compared to spoken numbers or numbers written as fifty dollars. In particular, we feel that the Arabic notation has a strong bias against an hyperbolic interpretation.

We agree that the Arabic notation might have a bias towards precise interpretations, as reflected in the optimal cost parameter. Regardless of this potential bias with the Arabic notation, participants still reported hyperbolic and fuzzy interpretations in ways that seem natural and consistent with the model. Different presentation formats could differ, and would be accommodated in the model by the cost mechanism.

It is worth noting that presenting the stimuli in spoken form introduces a great deal of complexity and potential confounds such as prosodic information, which we wanted to eliminate at least in the first step towards examining hyperbole understanding. We could try presenting the numbers written out as "fifty dollars," We could try presenting the numbers written out as "fifty dollars," which could remove this bias but perhaps introduce others. For example, since numbers are less frequently presented in words, by selecting this unusual presentation, we may unwittingly prime participants to believe that minimizing cost is not an issue in this particular experimental context, since writing the numbers out in words is clearly more costly than presenting the numerals. As a result, we decided that it was preferable to use the more common form and present the numbers in Arabic notation.