Keith J. Holyoak, Editor   
Department of Psychology   
University of California, Los Angeles   
405 Hilgard Avenue   
Los Angeles, CA 90095-1563

July 5, 2019

Dear Dr. Holyoak,

We would like to thank you, Emiel Krahmer, and the anonymous reviewers for the very helpful comments on our paper, “When redundancy is rational: A Bayesian approach to 'overinformative' referring expressions.” We believe that the reviewers’ comments, criticisms, and suggestions have led to a better and clearer paper. You specifically asked us to focus on providing a better explanation of the notion of a “continuous semantics” as well as toning down the theoretical implications surrounding it. We have now done so (see pp. XXX). We have also updated the Author Note as per your request.

The main differences to the previous version are:

1. XXX
2. XXX

Below are our point-by-point responses to the reviewers. We hope that you will like the new version of the paper. Please let us know if you require additional information. We look forward to hearing from you!

Sincerely,

Judith Degen  
Robert X.D. Hawkins  
Caroline Graf  
Elisa Kreiss  
Noah D. Goodman

# Responses to Reviewers’ Comments

**Editor**

[…] One basic issue (raised particularly by Reviewer 3) is to clarify the notion of "continuous semantics", and more generally to justify (and/or tone down) some of the general claims about implications for theories of meaning.

We now do so on pp. XXX.  
  
[…] As part of your Author Note, please provide the details (2-4 sentences) of prior dissemination of the ideas and data appearing in the manuscript (e.g., if some or all of the data and ideas in the manuscript were presented at a conference or meeting posted on a listserv, shared on a website, etc.).

We have now done so.  
  
  
**Reviewer #1:**

When redundancy is rational: A Bayesian approach to 'overinformative' referring expressions - Judith Degen, Robert X.D. Hawkins, Caroline Graf, Elisa Kreiss and Noah D. Goodman  
  
Reviewed by Emiel Krahmer (Tilburg University)  
  
This is an excellent paper, presenting a detailed, unified analysis of several distinct phenomena related to "overspecified" referring expressions. For this, the authors develop an adapted version of the Rational Speech Act (RSA) model, relaxing the assumption of deterministic meaning in favor of a graded semantics.  
  
There is lot to like about this work: the new theory is very interesting, and the analysis of the various phenomena is compelling (some comments below). Moreover, the empirical work is very well executed and the writing is excellent. All materials, data and scripts are publicly available on Github, which is good practice as well.  
  
Below, are some general comments, followed by some more specific, mostly minor points, none of which should pose a problem for the authors:  
  
- One phenomenon that is analysed in this paper is the asymmetry in redundant use of color and size adjectives. Various authors have explained this asymmetry in terms of incrementality. The current authors discard this explanation by pointing out that "(...) this would predict that speakers routinely should produce expressions like the blue small pin, which violate the preference for size adjectives to occur before color adjectives in English" (p. 7 - the same point is made on p. 63). However, I feel that this deserves more discussion, for a number of reasons. First of all, it would be good to make an explicit distinction between incrementality during content selection and incrementality during linguistic realization. PRO, for example, is explicitly developed as a model for conceptualization, and various other (also computational) theories assume that language production involves multiple stages. Levelt (1989, Speaking) offers an interesting discussion of 'discrepancies' between these different levels or modules, pointing out that it is perfectly possible for one concept or lemma to be selected before another, but nevertheless to be realized as an (audible) word later. Second, an analysis of speaking errors could also be interesting in this case. For example, it is plausible that corrections like "the blue uh small blue pin" are more likely than corrections like "the small uh blue small pin" (in fact, this is something that we found in our data, and this could easily be explained by an incremental strategy). Finally, in languages other than English (which have a less strict word order) it is conceivable that concepts are more likely to be mentioned in the order in which they are selected. In fact, Fukumura (2018, Ordering adjectives in referential communication, JML) offers some evidence for this. All this requires some more discussion.

XXX  
  
- Somewhat relatedly: you correctly point out on p. 11 that many authors replicated the preference of color over size, but this may also be due to the way color and size are represented in this earlier work. In a recent study (Van Gompel, Gatt, Krahmer, & Van Deemter, 2014) we made the size contrast extremely noticeable (using 'huge' and 'really small' differently colored objects) -- in that experiment we no longer found a preference for color. This might be worth pointing out.

XXX  
  
- The authors discuss the PRO model, referring to Gatt et al. 2013. It would be better to refer to the version published in Psychological Review in 2019. From this full version, it should hopefully be clear that the PRO model is not "specifically designed to capture the color-size asymmetry" (p. 7). In fact, that paper describes multiple experiments, also one involving more than 2 properties (and offer a generalization of arbitrary numbers of properties), not only color and size (so, it is not true that "it is neither flexible enough to be extended straightforwardly to other modifiers beyond color and size" p. 8).

XXX  
  
- Even though I don't feel this is necessary for the paper, it would be interesting (and useful) to compare the cs-RSA and PRO models in a bit more detail. One obvious strength of the RSA model is that it offers a more general communicative framework, and also has an account of the listener (even though this does not play a large role in the current paper). On the other hand, the PRO model relies on less parameters than cs-RSA (two, and even for those two it received criticism).

XXX  
  
- Relatedly, the best performing cs-RSA models rely on 4 or 5 parameters (Table 7), which reminded me of the famous quote attributed to von Neumann ("With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.") I wonder what your take is on this?

XXX  
  
- A crucial component of the new cs-RSA model is its reliance on a continuous semantics. While section 6.4 offers an interesting discussion of how the semantic values are determined, I would like to hear a bit more about the underlying intuition of this kind of semantics. Is there independent evidence for the use of a continuous semantics? And does it have implications for other phenomena that have been analysed using RSA?

XXX  
  
- Just a comment about audience design: we have long been thinking about an experiment in which speakers produce references for color blind addressees, but never came around to actually doing it. Given that there are different kinds of color blindness (e.g., red-green, or blue-yellow), it would be really interesting to see how a speakers develops strategies for specific color use from one color blind addressee to the next. The RSA framework would be eminently suitable to model this.

XXX  
  
- Finally, the text is an easy read, but perhaps also a bit long, and could probably be compressed a bit, for example, by removing some of the redundancies (e.g., it is mentioned multiple times that "some key patterns of overmodification" are reviewed in section 2; footnote 23 and 26 are essentially the same (not that removing one will save a lot of space), etc.).

XXX  
  
More minor comments:  
  
- p. 3: "For instance, speakers are likely to produce referring expressions like the small blue pin instead of the small pin in contexts like Figure 1a, even though the color modifier provides no additional information" ==> I am not sure whether earlier authors studying "overspecification" would agree with this characterization. The term is an unfortunate one (also see below); in this example, the color 'blue' is not strictly speaking needed for identification, but it does provide information (color), which can be helpful (e.g., for identification). Note that this paper sometimes also uses overspecification in this sense (e.g., on p. 34-35).

XXX  
  
- p. 4: "E.H. Rosch" and "E. Rosch" are the same person (I assume), so the use of the initials here is a small BibTex mistake (one reference with E.H and one with E.). Perhaps more importantly: Dale & Reiter do discuss basic level values (inspired by the work of Rosch), so it might be worth mentioning this here?

XXX  
  
- p. 15: The use of cost functions is very interesting. In an early study on generating referring expressions, we used cost functions as well, showing that different ways of assigning costs resulted in different models known from the literature. We also discuss the monotonicity constraint mentioned on p. 15 (Krahmer et al. 2003, Computational Linguistics).

XXX  
  
- p. 18: "Color adjectives are typically treated as absolute adjectives while size adjectives are inherently relative (Kennedy & McNally, 2005)." Here, I think, a reference to Pechmann would also be relevant.

XXX

- p. 19: "From this perspective, these redundant modifiers are not overinformative; they are rationally redundant, or sufficiently informative given the needs of the listener." I wonder whether it is helpful to introduce new terminology like "rationally redundant", given the amount of confusion that "overspecification" has given rise to.

XXX  
  
- p. 24: Both speaker and addressee received feedback for 1 second before the next trial -- maybe briefly say why this is done?

XXX  
  
- p. 25/6: "Articles were omitted in 88.6 % of cases and nouns were omitted in 71.6 % of cases. We did not analyze this any further." Still, I do wonder how you would explain that speakers mention nouns in 28.4% of the cases? Especially, because in the third study you do analyse the production of nouns.

XXX  
  
- p. 33 and elsewhere: this really is a detail but I wonder whether it would be good to brielfy say how you determined what to pay the Mechanical Turk workers.

XXX  
  
- p 35: it is interesting to observe that in study 2 10 speakers used "roundabout descriptions", and none did in study 1. Any idea why this might be the case?

XXX  
  
- p 42: It was a bit surprising how cost functions are defined here (in terms of length and frequency). Why didn't you do this in the earlier studies? Why not use the same cost function(s) throughout? Or maybe these more complex cost functions are just not needed (cf. Table 6 and the discussion on p. 45)?

XXX  
  
- p. 48: "Overall, super level mentions are highly dispreferred (< 2%), so we focus in this section only on predictors of sub over basic level mentions." The fact that super level mentions occur so infrequently is not surprising, but I do wonder how cs-RSA would account for this. Maybe you can briefly hint at this? (Incidentally, Figure 21 does seem to include super level mentions.)

XXX  
  
- p. 50: "Length was coded as the ratio of the sub to the basic level's length." Why is length computed here differently than in the previous study (number of characters). Would it make a difference?

XXX  
  
- p. 53: Figures 16a and 16b are easy to compare visually, would it be possible to do something similar for this study?

XXX  
  
- p. 55: "The reason that redundant referring expressions are interesting for psycholinguists to study is that they seem to constitute a clear violation of rational theories of language production." Here (or elsewhere) it might also be nice to mention cases, briefly discussed by Dale & Reiter (1995) like "sit by the newly painted table" in a situation where there is only 1 table -- this seems like a textbook Gricean example (the speakers wants to warn the addressee of the possibility of getting fresh paint on their clothes).

XXX  
  
- p. 74ff: There are a few minor bibtex mistakes in the references (e.g., 'only moderately Gricean ==> `only moderately Gricean; Animal , dog , or dalmatian ? ==> Animal, dog, or dalmatian?; something seems wrong with the doi in Maes et al. 2004; the reference for Mitchell (2013) is incomplete).

XXX  
  
  
  
**Reviewer #2:**

This paper presents an extension to the Rational Speech Act (RSA) model that allows for a probabilistic mapping between words and meanings (permitting some flexibility/doubt in how `blue' maps to certain colors). This extension has the advantage of allowing for the correct prediction of behavior in 4 domains related to reference. The 4 domains involve redundancy in choice of referring expressions -- asymmetry in overmodification patterns, increase in overmodification in complex scenes, increase in overmodification for atypical features, and the preference for basic level nominal reference. I'd be happy to see this published since I found it insightful to read and helped me see the connection between these phenomena. The paper already does more than most papers by unifying multiple phenomena. It raises new directions for future work and that's a good thing. I do have a couple questions/ideas below but I do not think they should preclude publication. Perhaps my biggest caveat is that I don't work with these models myself and so I'm not in a position to critique the implementation choices and presentation.  
  
My two biggest questions (and I don't think they're huge) are about Shannon's noisy channel model and about the intuition behind deterministic / continuous semantics. I'd always (perhaps naively) thought of redundancy as a strategy against message loss. So in terms of use of bits of signal to convey a message across a noisy channel, there's some tradeoff between message cost and ensuring that the message gets through. In that way, speakers might want to be (selfishly?) redundant in the way they package their message to avoid having to repeat themselves if the message doesn't go through. I couldn't tell if what RSA (or cs-RSA) is doing captures this intuition about the cost of having to repeat yourself (as a justification for preferring slightly redundant messages). Messages themselves have cost but a redundantly specified message seems to go some ways towards avoiding the cost of having to repeat yourself. Even if this intuition is relevant (not sure that it is), I don't think that it specifically predicts the phenomena the authors aim to account for -- e.g., would a redundancy-to-avoid-repetition strategy specifically predict a color/size asymmetry? Though maybe it is relevant in that it helps motivate why a speaker would be redundant beyond the one-shot communication that I think the models here are being tested on (even though the communication games that the humans played were multi-turn, if I understood correctly). The advantage of cs-RSA seems to be in showing why uncertainty in what is meant by a word can provide an umbrella explanation for a number of phenomena. So my question may be a misplaced one since I think you do adequately (and insightfully) address those phenomena. But I wondered if you could say a bit about whether that intuition of redundancy-to-avoid-repetition has a place in this model (or is already inherently built into it)? For example, p.15 notes that "the cognitive plausibility of complex utterances being cheaper than simple utterances is highly dubious", but not having to repeat yourself is also a cost-effective measure, no?

XXX  
  
The other question was a request to provide an intuitive example earlier in the paper. As is, the phrase "deterministic meaning" is used early on but it isn't until the model is presented (definition of `deterministic' on p.16) that the reader gets a glimpse of what a `graded semantics' will be. Even just a definition on the first page that deterministic meaning assumes that a word maps unambiguously onto its meaning -- that would help portray why one might need to relax that assumption (at p.12 the text still is asking for the reader's good faith that a definition is coming: "the most basic model, as formulated by M. C. Frank & Goodman, 2012, does not produce the phenomena outlined above due to its strong focus on speakers maximizing the informativeness of expressions under a deterministic semantics"). Maybe the top of p.6 would be a good place for an intuitive example where you'd point out (if I'm getting the picture right) that `blue' and `small' do not apply equally well to all roughly blue, roughly small objects, and that a speaker might opt to include more modifiers when any one alone might not be a perfectly apt descriptor, i.e., blueness and smallness are gradable and so hedge your bets and use both. If you can give an intuition for how a speaker, in knowing that `blue' and `small' are imprecise, would choose to overmodify with color more than size and more so when there's more variability in the scene, that would be great. On p.19 this is done somewhat: "The crucial mechanism is that when modifiers are relaxed, adding additional, 'stricter' modifiers adds information."

XXX  
  
  
Other notes in order:  
- I realize it's a bit silly to have printed the paper in black and white given that it's about \*color\* modification, among other things, but assuming some readers are like me with a B&W printer, there are a number of places where the figures are kind of meaningless without color and could be easily adjusted to be B&W friendly.  
- Figure1 & Figure3 - the text referring to fig1a/3a talks about the green border; just use a black/bold border?

XXX  
  
- Figure8: the utterance type for color/size/size\_color isn't distinctive in black and white and so all the dots look the same

XXX  
  
- Figure16: use dotted/dashed lines so the difference isn't reliant only on color?

XXX  
  
... Fig4 is clearly one that needs color so no real way to change that one; I guess Fig21 is also hard to do with no color (maybe use shape?)

XXX  
  
... Fig8 works fine with the use of shapes for 2/3/4 distractors  
  
- p.4: Does Grice talk about a 'good speaker' or a 'cooperative speaker'? I'd expect the latter since otherwise it sounds more prescriptive than I think Grice ever intended

XXX  
  
- p.5: "Figure 18"? Probably Fig3

XXX  
  
- p.16: Here or elsewhere, is it worth acknowledging more that assigning truth values for color properties is a slight variant on the more typical assignment of truth values for propositions? I mean, the text puts scare quotes around `true' for "a size adjective being `true' of an object" but I did have a moment's pause in thinking about the truth of the adjective, since it seems there's a slight distinction between a command "click on the blue" and the true/false observation "this object is blue". Probably overly philosophical/pedantic and not critical to the implementation, but just gave me pause.

XXX  
  
- p.20: typos - insifficient/coounts

XXX  
  
- p.21: "Scene variation is maximal when all distractors except for one (in order for the dimension to remain insufficient for establishing reference) are of a different color than the target."  
... does this assume an ability to easily assess the number of dimensions of variation - i.e., to inspect all items and get inventory of dimensions? Does this scale to real-world scenarios? This is also a critique of work by Gatt and others (going back to the incremental algorithm) that seems to assume it's easy/possible to enumerate all the dimensions in a scene so it's not specifically your problem.

XXX  
  
- p.24: The description of the procedure mentions that a red border appeared around the clicked object in both the listener and speaker's displays and then the next sentence says "That is, both participants received feedback about the speaker's intended referent and the listener's inference" but did the description specifically mention how the listener learned what the intended referent was?

XXX  
  
- p.25: either explain "deliberately" or omit?  
"We deliberately chose photo-realistic objects "

XXX  
  
- p.25, p.36, p.47: Given the length of the paper, I was trying to keep an eye out for things that could be shortened. These descriptions of the automated data pre-processing could definitely go. All the reader needs to know is the upshot -- that the expressions were annotated and how many were excluded. The details about the R script and the subsequent hand-checking aren't really relevant (that I can see) since the upshot is the same that everything got annotated carefully.

XXX  
  
- p.26: typo --  
"proportion of distractors that does not share"  
-> "proportion of distractors that do not share"

XXX  
  
- p.26: stats coding -- The later mixed effects model was described with the full details of how factors were coded (p.40 - "all predictors centered"). Could that be added here for the fixed effect of sufficient property color/size?

XXX  
  
- p.26-27, p.49: it's a bit odd that the pattern of results is reported first in a footnote and then in the main text (e.g. p.26 is footnote 18 about convergence and the alternative model and then the main text reports the results on p.27). Move footnote?

XXX  
  
- p.34: Was the abbreviation "cc" defined (I figured it out as color-competitor but it was briefly unclear)

XXX  
  
- p.35: did you mean "expressions that included \*superordinate\* categories ("yellow fruit")"?

XXX  
  
- p.40: how neat that speakers sometimes used 'banana' to refer to a yellow banana even when other bananas were present!

XXX  
  
- p.40: I was surprised by this: "a main effect of color competitor presence, such that color mention was more likely when a color competitor was absent"... really? more color mentions when there was no color competitor?

XXX  
  
- p.45: interesting the finding and replication that cost didn't have an effect! seems to open up some channels for future work.

XXX  
  
- p.48: The text says that there were "41 trials on which the listener selected the wrong referent were excluded, leading to the elimination of 2.1% of trials"  
... but then farther down the page it says " The clearest pattern of note is that sub level mentions are only preferred in the most constrained context that necessitates the sub level mention for unique reference (e.g., target: dalmatian, distractor: greyhound; see Figure 18a). Nevertheless, even in these contexts there is a non-negligible proportion of basic level mentions (28%).'

XXX  
  
... so in the 28% of cases where a speaker said "dog" in a context with 2 same-category objects (dalmatian & greyhound), how did the listener get the right answer? Were those cases like "banana" mentioned above where the term "dog" could successfully refer to the more prototypical dog? But is that enough to yield only 2% error overall?!

XXX  
  
- p.48: for the results, it looks like a logistic regression was used to predict sub over basic level mention, but earlier on the page it mentions 1872 cases classified as sub/basic/super. Were the super cases eliminated for the analysis? p.51 mentions some patterns of use of the super term so I was getting confused how the regression was constructed.

XXX  
  
- still keeping an eye out for ways to shorten the paper since it does feel rather long (or else a more clear justification of what misunderstanding/conclusion the inclusion of this helps avert):  
- p.54 - comparison of model components across experiment  
- p.56 - speculation with stapler example  
- p.59-60 - continuous semantics (I found myself skimming at this point since I'd been convinced by the advancement that cs-RSA offers and it was already page 60 and it felt like a section that a previous reviewer had asked the authors to add)  
- p.63 - footnote about the interaction at CUNY with Sarah Brown-Schmidt

XXX  
  
- p.59: typo 'compoenent'

XXX  
  
- p.62: the portrayal of the debate between audience-design vs production-internal processes brings me back to the issue raised early on of whether a speaker can selfishly want to be clear in order to avoid having to repeat/reformulate what they say -- so selfish/altruistic goals go hand in hand if there is a preference (a selfish/altruistic one) for successful conversational interaction.

XXX  
  
- p.63: I had been wondering if cs-RSA would account for Rubio-Fernandez's 2016 finding that Spanish speakers show less overmodification. In principle, it doesn't seem obvious how it would since the order of the attributes in the referring expression isn't specifically controlled and Rubio-Fernandez argues (I believe) that the English preference to overspecify reflects the helpfulness afforded by a color word early in the utterance (helpfulness for visual search), whereas less benefit (for visual search) is provided by the color word once the disambiguating shape word has already been mentioned. The paper here gets at the issue of incrementality and attributes Rubio-Fernandez's results to English speakers' inclusion of redundant color adjectives to buy time for noun planning:  
"This predicts that in languages with post-nominal adjectives, where this delay strategy cannot be used for noun planning, there should be less redundant color mention; indeed, this is what Rubio-Fernandez (2016) shows for Spanish."  
  
... I guess I'd still wonder whether that noun planning benefit is enough to explain such a difference between Spanish and English. If speech planning is what's relevant, it almost seems to predict there should be some equivalence whereby English speakers can be said to be buying time to plan the noun by including the color word and Spanish speakers should be buying time to plan the adjective by including the noun word (but yet they don't bother to include the adjective!). In the data that Rubio-Fernandez presents, the colors and shapes get used over and over so I would imagine the words for them don't present great difficulty for lexical retrieval. It seems there's something else about color overmodification \*early\* in an utterance which is linked to improving success on the conversational task by helping the listener's visual search. I don't think cs-RSA needs to account for this currently (it already does a lot!) but the paper seems to slightly discount this interesting puzzle by casting it as an issue of incrementality and speech planning.  
  
XXX  
  
**Reviewer #3:**

When redundancy is rational: A Bayesian approach to 'overinformative' referring expressions  
  
Summary  
  
This paper presents three studies addressing the use of redundant or overinformative adjectives. The results are interpreted as lending support for an extension of the Rational Speech Act (RSA) model, so semantic truth values are drawn from a continuous interval [0,1] rather than the traditional binary range {0,1}.  
  
The central empirical focus of the paper centers on the observation that speakers are more likely to provide additional modification for color adjectives ("blue vase") compared to size adjective ("large vase"). Three Internet experiments are reported, all of which were conducted with pairs of subjects, in which the propensity of including an adjective in an "utterance" is measured. The experiments appeared to be well-conducted, and the analysis was fairly sound. The findings largely replicate the results of previous experiments, including the typicality effect (where subjects tend not to modify nouns with their typical color properties, e.g., yellow banana vs blue banana), as well as color size asymmetry, scene variation, and (sometimes) a length or frequency effect. As I understood it, the central novel finding was the interaction in which scene variation affected size adjectives more than color adjectives, along with the model-fitting of the data throughout.  
  
Much of the paper is admirably clear and well-written, and addresses an area of current interest within a framework that has steadily been gaining traction. However, there are some concerning issues regarding consistent and appropriate use of key terminology, methods and reporting, and the interpretation of the results. After pointing out more general issues, I detail my concerns in each of these areas. I believe these issues to be considerable, and so I cannot recommend this paper for publication at this time.  
  
  
General issues  
  
\* As it stands, the paper is much too long. The discussion is often clear but unnecessarily detailed and at time redundant with past sections. Specifically:  
o The final section 6 contains a good deal of useful discussion but should be tightened and shortened a great deal,  
o Much the experimental methods and data handling are essentially identical across experiments; their introduction could be compressed and streamlined,  
o Some irrelevant information is included, and discussion of past findings seem to drag on, and  
o Each section starts with very similar research questions, which feels repetitive and unnecessary.

XXX  
  
\* However, more technical issues are sometimes given too short of a treatment, and key terms are not introduced in the appropriate places - e.g., the synonymy between "noisy" and "continuous". Also, explaining crucial procedures in model fitting like normalization would be very useful. So, the paper assumes too much technical background on behalf of the reader, but then spends too much time explaining issues that don't come up consistently.  
  
XXX

Terminological issues  
  
\* Inconsistencies in terminology make paper unnecessarily hard to follow:  
o Previously mentioned "noisy" vs. "continuous" terminology  
o Misleading or less than ideal choice of terms: I would personally prefer "discrete" to "deterministic", as "discrete" applies to the finite (binary) choice of truth-value options, whereas "deterministic" applies to the process of selecting a value. The terms are not synonymous: e.g., one could have a discrete, but non-deterministic function.

XXX  
  
\* I am uncomfortable the general use of "utterance" as a single word or pair of words. This appears to be common in current RSA literature, and I believe it is linked to the idea that RSA is non-compositional (and so the internal composition is effectively ignored). Why redefine the standard use of utterance to apply here?  
\* Minor issue on p 26: "The model included the maximal random effects structure that allowed the model to converge: by-speaker and by-item random intercepts." This seems misleading. Maximal models refer, by convention, to by-subject and by-item random slopes and intercepts. While technically correct, the description doesn't

XXX  
  
  
Methods and reporting  
  
\* Given that the only method for data collection was an online game, not much is known about the subjects, their native language ability, or their strategies. Were there any controls for native language proficiency or for attention or cooperativity in the experiments? I do a good deal of online experiments, like everyone else, and find that controls to identify inattentive or uncooperative subjects is crucial, even with large sample sizes.

XXX

\* I would have liked to have known whether the directions that the "Speaker" gave were informative. That is, if the Speaker elected to provide "overinformative" modifiers, were the Listeners better able to perform the task? It's not clear to me whether this information can be extracted from the data, but it would be highly relevant with respect to the issue of "audience design", discussed throughout the paper. Similarly, it would be helpful to know if strategies of overmodification changed throughout the experiment, or if some participant pairs were more apt to use it than others.

XXX

\* In my mind, a reader should be able to replicate the essential design of the study and the components of the quantitative model. I do not think that the authors have provided sufficient information to permit replication.

XXX  
  
Interpretation  
  
\* In what sense is this really a semantic theory? I take no issue with RSA being non-compositional, provided it is a pragmatic theory of issuing speech acts, or some quantitative model that captures the extent to which an utterance (or string of words) can be issued, given a speaker's uncertainty about how to describe the world.

XXX

\* In general, the motivation for a continuous semantics is weak, and, frankly, a bit hyperbolic and overblown at times. Although a more measured discussion is found in the General Discussion section, the Introduction reads like a manifesto against "deterministic semantics", without fully recognizing that this decision involves rewriting all of semantic theory, without ever truly addressing the compromises that such an enterprise would have to take. I remain unconvinced that the paper sufficiently motivates such a decision, or that the machinery justifies the payoff for a relatively small domain (which has other, though perhaps not computationally implemented, solutions).

XXX

\* Instead, it seems entirely plausible that this model captures something about how a (rational) Speaker behaves, rather than the conventional meaning that a word has in its local context (roughly, the standard definition of a semantic theory). The same point could be made if the word "semantics" never appeared in the paper, and the model were recast in terms of use conditions on licensing utterances. It's a given that uncertainty about the world abounds; for example, speakers may not know (a) what the model of the world truly is (and so be unable to assess whether a term can be predicated of an object truthfully), or (b) what the common ground between interlocutors is, and, consequently, what the reference set for relative adjectives is. None of this need involve rewriting the traditional bivalent truth values in the semantics.  
  
XXX  
  
In sum, this works makes a valiant and commendable effort at providing a quantitative model for a troublesome domain within more traditional semantic theory. The central issue, for me, resides in the extent to which the fairly radical innovations are truly justified by the relatively small empirical gains made here, especially when it seems that similar or virtually identical points could be been made if the framework were cashed out over slightly different objects of interpretation (use conditions vs. meanings).

XXX