Dear Rick, Dear reviewers,

first of all, we would like to thank you very much for your close reading of our manuscript and the many helpful and constructive comments. We have tried to address all of the points raised in the best way we could, while also keeping an eye to the length of the paper. Below is a detailed list of how we responded to comments and suggestions.

The main changes in the paper are:

1. a shortened and tightened introduction, aimed at avoiding misunderstandings before crucial ideas and terminology has been introduced properly;
2. a complete make-over of the discussion of “core theories” *traditionalism* and *grammaticalism*, especially a completely rewritten Section 3 which introduces these; and
3. an added Section 6, which took over material from the now shortened Section 5.4, dedicated to a general critical discussion of our methods and the interpretation of our results in light with the reviewer comments, including a very brief description (in the interest of overall length) of a post-experiment that we ran to properly address the well-taken methodological calms of Reviewer 1.

Core sections 4 and 5 are mostly as before, but we have changed many formulations in wording and emphasis to make clear the role of prosodic manipulations, our preference-related controls, the Bayesian analysis and other issues that were brought up.

We hope that it is recognizable that we wanted to respect every critical point raised. We are fully aware that there are many outstanding issues in this difficult and controversial terrain. Our aim is to provide a constructive empirical step forward, without polarization, and we would be grateful for any further thoughts and comments you will surely have to help us make achieve this goal.

Thank you very much for your efforts!

Best wishes,

the authors.

## Editor’s comments

Both review 1 and 3 point out a worry regarding the experimental setup (or more accurately regarding the interpretation of certain responses). If I understand them correctly, they express the same worry: **the danger with a verification task is that subjects accept the first compatible state. If that's a possible strategy, then it is unclear how to interpret your data**.

Actually, the reviewers have expressed several doubts about our paradigm and the interpretation thereof. We have tried to address all of these without expanding the paper too much in a novel Section 6 (“Reflection”). The most important addition is a short description of an additional experiment that we ran to address Reviewer 1’s worry that our task might be insensitive to implicature readings altogether. See below for a detailed reaction to the more specific comments of the reviewers.

This is tricky. It is very hard to do justice to such a fresh (and still moving) theoretical landscape. Clearly, both the grammatical camp and the traditional camp host a number of various more specific theories. Reviewer 2 makes an important point connected to that regarding the theoretical significance of the experimental result. The "camps" make very coarse predictions. The predictions the paper alludes to are the result of very fine design choices, which quite many scholar on the relevant side of the divide would disagree with.

There's a choice here: **either you show that your results are relevant more generally, or you concede that things are much subtler and you sketch the consequences of your data regarding the debate much more precisely**.

The point is well taken. We have tried to do both. We reformulated every relevant part of the paper to emphasize even more clearly than before that there are alternative options. Section 3, which introduces the theoretical positions, has been completely rewritten and there is also the new discussion section 6, which addresses theoretical implications of our results more carefully and in more detail.

Throughout, we draw on Chemla & Singh (2014, Language & Linguistics Compass) and adopt their terminology: we would like to compare “core theories”, but only under additional “auxiliary assumptions” will the core theories yield falsifiable predictions about empirical data from a concrete experiment. (Chemla & Singh are concerned with processing models, but the same applies for predictions about behavioral data independent of processing (e.g., whether to accept a sentence as true or false).)

Section 3 gives additional motivation and justification for our focus on the versions of traditionalism and grammaticalism that we address at the outset. To accommodate Reviewer 3’s criticism, we especially motivate our focus on the Strongest Meaning Hypothesis. We agree that it is a moot point arguing about “prominence” of an idea in the literature, as this is by necessity a subjective notion. But the SMH has been used in other domains as a disambiguation criterion, and is certainly itself an established, well-known notion, which means that our results are more generally relevant as well, beyond the debate about embedded implicatures.

But of course, at the end of the day, we totally agree that there are many more variations on “traditionalism” and “grammaticalism” that could be maintained, and therefore we consider more alternatives, as suggested by the reviewers, in Section 6. We stress, in the paper and here, that nothing we say decides between core theories for good, but that our data are nevertheless relevant in guiding the choice of “auxiliary assumptions” for either core theory. We believe that this is as much as can be achieved given the scattered theoretical landscape and controversy about how to properly map theoretical positions to empirical data. Still, we believe that ours is a constructive contribution that speaks exactly to this theoretical and methodological uncertainty by giving a study that tries to disentangle data on alleged implicature readings from confounds due to silent prosody and pictorial effects.

## Reviewer 1

I do have one minor concern about the theoretical review. **The discussion of typicality accounts of some previous data needs elaboration**. I'd like to know what the authors think is typical of what, and why. What's typical about having some (or all) of some squares connected to some (or all) of some circles?

We try to elaborate on what “typicality” means in footnote 6. The reason we are hesitant to venture into a lengthy discussion, is because the matter is very controversial, especially when it comes to the theoretical interpretation of “typicality effects”. It is quite intuitive that there are such effects, and this is all that really matters for our argument that we should be careful in our design to avoid potential “pictorial confounds” as much as possible.

I have one major qualm about the task. **Everyone agrees that "some" is typically strengthened in a simple sentence.** **Imagine what the present incremental verification task would be like with a simple sentence.** A subject would hear something like "The letter is connected to some of its triangles." The subject then sees a picture in which one letter is shown connected to three triangles, with the connections to some unseen triangles obscured. Would the subject typically say "true" of this picture? I bet that many, perhaps most, subjects would. This would be a "literal" reading in the terms of the present paper. Would the authors thus conclude that scalar terms are not strengthened in simple sentences? To be sure, they could do an experiment like this (ideally including such simple sentences with the range of sentences studied in the present experiment) and show that I lose my bet. But **in the absence of such evidence, I cannot take the present data to say anything specific about embedded implicatures**.

This is a very serious and very good point, and we are very grateful that is has been brought up. The only proper way of addressing this issue is to bite the bullet and run another test. This is what we have done. We implemented a short version of the incremental verification task, whose critical items were “simple” scalar implicature sentences with *some*. The results are clear: the task is generally sensitive to implicature readings; almost one third of the answer are “implicature answers”, even if these can be given only later than a literal answer.

It could be objected that there are too little pragmatic answers, showing that the task is still too insensitive. But that would not be fair. For one, the rates of implicature vs literal answer vary from task to task. In a simple picture verification task there are also only about 35-40% or pragmatic answers (e.g., Benz & Gotzner 2014). Also, it has recently been suggested based on corpus-data (Judith Degen, to appear in *Semantics & Pragmatics*) that non-epistmemic scalar implicatures are by far less frequent than commonly believed. Be that as it may, we hope that the fact that “only” almost one third of the responses were pragmatic is convincing enough to show that in principle our task is sensitive to implicature readings.

1. **conflation of pitch accent and intonational phrase boundaries under term “intonational” manipulation**
   * + - * check wording
2. **concern about “preference-related controls”**

The authors assume that there is a strong late closure effect for post-nominal prepositional phrases ("This letter is connected with circles and squares with suns"). **There is a huge "relative clause attachment" literature on items like these** (both relative clauses and prepositional phrases) that is often interpreted as evidence against late closure preferences in that some languages show a high attachment preference. German is one such language, for relative clauses, if little or no preference for prepositional phrases (e.g. Hemforth, Konie3czn, & Scheepers, 2000). Sometimes, however, the interpretation is that lots of factors matter (e.g., Gilboy et al., Cognition 1995). Anyway, it's an odd 'control' even though there is a pretty strong preference for late closure in the data, especially strong when the sequence of images requires that an early closure interpretation be manifested as a "false" response early in the sequence of images. **The effect of prosody here is clear but not particularly novel (and could be buttressed by citations of previous reports that prosodic phrase boundaries affect attachment of postnominal modifiers)**. The success in finding such an effect here but not in the scalar sentences can't be securely interpreted, given that the prosodic manipulations were totally different (pitch accent vs. intonational phrase boundary).

* + - * + address the relevant mentioned literature; add references for prosodic effects [PETRA]
        + stress that it is not a problem for us, because our analysis can deal with any type of preference over readings
        + acknowledge in discussion section that different types of “prosodic / intonational manipulations” could have had different effects [we simply cannot rule this out; the comparison is close to the best one can do] [MICHAEL]

1. **statistical analyses [MICHAEL]**

There were some parts I did not understand (e.g. p 30, comparison of models with 3- vs 2-level Reading factor, **why does the "more complex" model have fewer df than the less complex?**).

Since this is a comparison of nested models, we followed the convention of reporting residual degrees of freedom, i.e., the degrees of freedom of the residuals, which is inversely related to the degrees of freedom of the models.

Further, **the "generative Bayesian model" strikes me as egregious overkill**. The basic results from the main experiment are terribly simple: literal readings are dominant, with local readings making occasional appearances. The results from the "preference controls" are not as simple, but all I see the model doing is showing that there is a positional/sequential bias that messes up the results for these controls.

We have tried to work out the significance of the Bayesian model more clearly in the text. To repeat it here, the Bayesian model is important to address the question of how much a potential bias towards answering early or late should affect our judgment whether there are local readings at all. Classical statistical methods simply cannot deliver here, because they are built on models that are not specific to the incremental task, i.e., do not take into account the sequential nature of the task. We agree that the analysis does not contradict what a commonsense inspection of the raw data suggests. But that does not mean that the analysis is superfluous, does it? In fact, without the model other readers would certainly demand a sound test that addresses the question whether the “occasional appearance” of local readings in AS-sentences is significantly different from error responses.

In order to further put the model into perspective, we also reflect on it in Section 6.

The main experiment is ingenious and potentially very informative, \*if\* it is true that listeners do assign a single determinate interpretation to an AS or ES sentence (**my worry is that they may merely leave the possibilities open, to be settled by other information that may be or become available**).

We apologize for this misunderstanding and are very grateful for this comment, because it shows that the idea and motivation behind the Bayesian model has not been stated clearly enough. We address exactly this worry in the discussion on page 44. In short, the Bayesian model does not assume that subjects fix an interpretation in advance, but that they do assess interpretations at every step.

## Reviewer 2

My main concern, which is what I’ll largely focus on in this review, is with the assumed theoretical background and therefore also with the theoretical consequences the authors draw from their data. In particular, **the main claim – that the data are problematic for both ‘traditionalism’ (an inapproriate name, I think – more on this in (3) in section 3 below) and ‘grammaticalism’ is exaggerated, at best. So far as I can see, most of the data are easily accommodated in both theories without much ado.**

We are very grateful for making us see that our initial assessment was unsatisfactory. The new version of the paper tries to fully acknowledge the reviewer’s well-taken points by (i) fully rephrasing the exposition of “core theories”, (ii) more rigorously motivating the choices of “auxiliary assumptions” necessary to derive non-trivial *ex ante* predictions, (iii) stressing throughout which of our conclusions apply to the “core theories” and which apply only to the “core theories” + “auxiliary assumptions” that we fixed for the sake of exposition, and (iv) a more general reconsideration of other variants of the “core theories” in Section 6, including the very interesting suggestion made by the Reviewer about Magri’s proposal.

**There is no justification for the authors’ decision to use the Strongest Meaning Hypothesis (SMH) as representative of grammaticalism**. The authors say that it is ‘the most prominent’ of selection principles, but this is not a good justification.

We agree with the reviewer that additional motivation is necessary, and are grateful for pointing this out clearly to us. But we also believe that there can be a justification for placing the SMH into focus. As we argue in Section 3, the version of strength-based disambiguation that has been spelled out by Chierchia, Fox and Spector (2012, Semantics Handbook) makes clear and non-trivial predictions about reading preferences in an experiment such as ours where there is no clear question under discussion and where it is moot to speculate which formal alternatives are relevant. Moreover, the SMH is an established notion that is called upon in other domains for disambiguation as well. We hope that this does give us some justification for focusing on the SMH, in order to enter the experimental realm with something concrete to test, especially since we reevaluate “grammaticalism” as a “core theory” in the light of our data and are clear that we do not wish to claim that our data could refute the whole “core theory” as such.

The force of this objection gets its bite when we see what happens if we consider what is predicted under a different approach, say Magri (2009, 2011). Unless I’m mistaken, the approach actually makes quite a lot of sense of the authors’ data once **we assume with Noveck and Posada (2003) (among others) that participants can often be classified into (i) those participants that generally select the literal meaning, and (ii) those participants that generally strengthen**.

Just to address this point directly here, so as to avoid any further misunderstanding. There was no indication *in our data* that subjects consistently answered with a literal or a local reading. Furthermore, the idea that there are “subjective types” of answerers is, to the best of our knowledge, sporadically floated, but not stringently addressed. For what it’s worth, we mention this here as further justification for our position: we were very interested in this issue as well, and ran a MechTurk pilot of a picture-verification task in which we presented pictures like in Geurts & Pouscoulous experiments (but slightly clearer arranged to compensate for Chemla & Spector’s worries). We recorded judgments of each subject for both ES-sentences and simple unembedded implicature sentences. We were interested in whether there was a correlation between how each subject would rate unembedded implicature sentences and ES-sentences (e.g., those who respond pragmatically in an unembedded context would most likely also choose a “local answer” for ES sentences). The result was disappointingly clear: there was no hint of a correlation at all. This does not refute the reviewers point below, but, from our point of view, undermines the motivation for treating the idea as a principled *ex ante* account.

**4. Magri’s theory as an alternative variant of grammaticalism**

after (rightly) rejecting the SMH selection method, the authors are too quick to dismiss grammaticalism itself; I’d encourage them to clarify the nature of the challenge that remains for grammaticalism if it is to account for their data

* + - * + address Magri in the discussion section as another way out for grammaticalism
        + but note that Magri does not make any *ex ante* predictions and needs a linking function as well

**5. ’default assumptions’ of ‘strong traditionalism’**

the authors establish a ‘weak traditionalism’ and a ‘strong traditionalism’ based on whether, by default, crucial contextual features (e.g., speaker-opinionatedness, etc.) are either assumed to hold or not hold. But are non-default strategies not conceivable? […] It might be that unless these features are clearly specified in the context, the listener will need to make a guess about whether the speaker is opinionated, say, and different people might make different guesses based on all sorts of factors.

* + - * + misunderstanding of the term ‘default’ here
        + what the reviewer describes is exactly what we have in mind as well
        + whether our subjects make the necessary auxiliary assumptions or not, we cannot say for the experiment at hand; we therefore consider two versions of “traditionalism”, simply because we have to leave this open
        + change this in the exposition of ‘traditionalism’

**6. unjust to challenge traditionalism for failure of PMH**

More problematic, however, is the authors’ suggestion that a challenge for traditionalism is to explain their finding that accent does not increase the availability of local readings.

* + - * + I don’t see where we make any such suggestion
        + read and check!?!?

**7. terminological qualms about “traditionalism”**

I don’t see how ‘unrestricted traditionalism’ (p.8) is a kind of traditionalism. The alternative in (8) does not even entail AS. Why, under a traditional Gricean view, should it even be considered as an alternative?

* + - * + it simply needs to be addressed that the global reasoning scheme of Grice COULD explain local readings for AS-sentences (e.g., Chemla & Spector), so we subsume this near possibility under the term “traditionalism”

**8. grammatical approach not a theory**

* + - * + make clearer that we mean “without a linking function”

**9. difference between EC/LC ambiguity and implicature**

* + - * + acknowledge in discussion
        + still, if people favor the literal reading because that one is easy and they are lazy, it still holds that they favor the literal reading, and that needs to be explained (by some linking function)
        + in fact, it is exactly what our “economy principle” suggests (the one that the reviewer did not see sufficient motivation for; argh!)

**10. term traditionalism is ill-chosen**

* + - * + okay, but why so prickly?
        + could use “globalism” or just include a disclaimer to make any strong commitment to what tradition does and does not entail
        + I don’t think that the issue is about domain-specifcity and domain-generality of the strengthening mechanism

### Reviewer 3

1. **general worry about “early exits”**

there is a concern that participant may have responded as soon as a reading was compatible with the state of affairs regardless of whether that reading was indeed the preferred reading.

* + - * + this worry is what the preference-related controls are for: this case shows that it is not **generally** so that participants exit at the first possible reading
        + it might be that the type of ambiguity/underspecificity in AS- and ES-sentences is so that subjects prefer to exit at the first possible position
        + but that, too, is disconfirmed already by the data, in particular the ES-sentences AND the LOCAL responses in the AS-condition, which is not plausibly categorized as ERROR responses
        + be careful to work this out clearly in the discussion

**2. enlarge on discussion of potential confounds and/or alternative interpretations**

An elaboration on the possible confounds that this new and useful paradigm may introduce is warranted and is blatantly missing in the discussion of the results of the experiment and the general discussion.

* + - * + add to the discussion

**3. alternative explanation No. 1: different type of ambiguity EC/LC vs. implicature**

A possible explanation for the difference between the controls and critical items is that when participants are requested to choose between two readings of a structurally ambiguous sentence, they need to decide between two distinctive LFs or propositions, whereas in the case of the critical items, participants are required to distinguish between different parses, only one of which is associated with the literal reading of the sentence.

* + - * + add to the discussion
        + if phrased as difference in terms of processing (early vs. late or conscious vs. late disambiguation), then this makes total sense, but would be beyond the scope of this behavioral study;
        + the contribution of this study could be to highlight these potential differences and to pave the way for subsequent processing-oriented studies that can address these worries
        + acknowledge difference between EC/LC and implicature also when introducing conditions
        + possible connect to Reviewer 1’s worry about the different types of intonation used to favor different readings

**4. alternative explanation No. 2:**

An alternative interpretation is also possible, however: Under a surface reading of the sentence, the pictorial information necessary to evaluate *Exactly one bell is connected to some of its semicircles*, participants wanted to wait till two bells are revealed in order to know whether more than one bell is connected to some of its semicircles, independently of the fact that the global reading became false after step 2.(…) In sum, participant’s hypothesized goal to check that *exactly n* is true may mask their preferred reading, as the pictorial step that sheds light on the former also corresponds with the literal meaning.

* + - * + we do not understand this alternative explanation
        + what does it mean to “check that *exactly n* is true” without paying any mind to what property these exactly n objects are supposed to have?
        + what is the theoretical motivation for this alternative explanation?

**5. QUD for our task**

How about the idea that given the task, the question (not phrased as an open proposition, obviously) is “give me the picture that best matches the sentences,” with a bias toward the first unambiguous reading that matches the picture.

* + - * + interesting proposal, but how to motivate it theoretically and ex ante?
        + don’t address this at all!?

**6. “quick on the trigger” responses**

I wonder if the authors should be more cautious about incorporating the any observations from the surprising true responses on step 2 of *exactly one...some* into any semantic-pragmatic account for implicature. There’s always the possibility that these cases of “quick on the trigger” responses are not guided by participants’ deciding on a unique vs. non-unique interpretation of exactly. A similar type of early response occurred in the preference-related controls (Table 5) as well, even though an early response might turn out to be incorrect.

* + - * + we are not incorporating these observations into a new semantic-pragmatic account of implicature, but suggesting that there might be problems for experimentally validating any given such theories, based on unanticipated readings of modified numerals
        + it is not clear to us which responses the reviewer has in mind for the preference-related controls, as there are no similarly high numbers of errors that cannot be explained as spill-overs or true-false errors

**7. provide a uniform set of modifications to account for AS- and ES-data**

The modifications the authors seems incongruous. Could the authors suggest modifications for each view that would account for both the AS and ES accounts?

* + - * + we certainly would, if we could; but we don’t see how (yet)
        + unfortunately, we must pass this over as an open problem that we cannot fix

### Main ToDo’s

* highlight the need of linking functions to turn theoretical positions into predictive theories testable in the lab (cite Chemla & Singh)
* consider our contribution as a theoretical investigation of some (hopefully plausible) linking functions, and a suggestion of an experimental method for testing reading preferences

### Questions for Rick

1. Reviewer 1: “leave it for psychology … not JoS” -> but that’s what we do, so not worth it?
2. what to make of Reviewer 3’s second alternative explanation suggestion