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 of 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 formulations 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 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. 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 our 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 2’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 for our argument, we think the essential consequence for our argument is 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 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 answers vary from task to task. In a simple picture verification task there are also only about 35-40% of 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. 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.

Apart from this major concern (and a concern about the "late closure" items, below) I have to say that the experiment appears to have been very carefully thought out and very carefully conducted. The use of both AS and ES sentences is admirable (despite the apparent strange interpretations a few subjects made of "exactly one"), the prosody manipulations are interesting (despite the rather extreme pitch movement in the AS/ES sentences, and the conflation of pitch accent in these sentences and intonational phrase boundaries in the "preference" sentences as an "intonational" manipulation), and the existence of the positional control sentences is admirable.

Thank you for pointing out this issue. In order to avoid terminological confusion, we now subsumed prosodic phrase boundaries and pitch accents under the more neutral notion of a “prosodic manipulation”.

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).

Thank you for giving us to chance to clarify this point. We now formulated the respective section a bit more carefully and updated the references. In the following, we will briefly elaborate on this issue.

First, we agree with Reviewer 1 that the Late Closure principle might not exclusively account for final interpretive preferences of all kinds of adjunct attachment ambiguities. Indeed, relative clause attachment preferences may vary cross-linguistically or even within one and the same language. For instance, though a high attachment preference has sometimes been reported for German (e.g. Hemforth et al., 1998), other studies observed a low attachment preference across a range of different experimental tasks (e.g. Augurzky, 2005). Different suggestions have been made to account for these discrepancies, such as the thematicity of the items involved, implicit prosody, or the availability of different constructions (e.g. Frazier & Clifton, 1996, Grillo, 2014, Fodor, 2001). To discuss these proposals would go beyond the scope of the present paper.

For PP attachment ambiguities involving two nominal hosts[[1]](#footnote-1), the majority of previous evidence suggests a uniform and probably universal preference towards low attachment (see Fodor, 1998, for a brief overview; and Clifton, Carlson & Frazier, for evidence on highly comparable structures). Crucially, a low-attachment preference for PPs (60%) has also been attested for German (e.g. Hemforth et al., 2000, see (1)).

(1) Die Tochter der Lehrerin aus Deutschland traf John.

The daughter the teacher from Germany met John.

“The daughter of the teacher from Germany met John.”

This effect appears to be stable across Experiments (Questionnaires 1 and 5), but can be affected by pragmatic anomalies (Questionnaires 2 and 3, where an NP2 decision would have resulted in an implausible reading), or by prosody (auditory Questionnaire 5, in which a prosodic bias towards high attachment had been (accidentally) realized). In our study, we explicitly manipulated prosodic phrasing and found clear-cut preferences, which are modulated by our prosodic manipulation. Importantly, there is no pragmatic bias in the data that might have made one or more of the available readings more likely. We hope that this addresses the reviewer’s worry, but if there are other studies or if we misread the mentioned paper, we are happy to be told and reformulate accordingly.

In any case, we would like to stress that even if there was no preference in German or the reverse preference, this would not be fatal to our general setup. In fact, our design and our analysis does not strictly depend on whether preference-related controls show exactly the bias that we hypothesized that they would show, given the literature and our own intuition.

Also, that the kind of prosodic manipulations were different in preference-related controls and critical conditions is true, but does not critically affect our main argument. The prosodic markedness hypothesis is a way for traditionalists to allow for local readings *if* they are prosodically marked. The hypothesis is not that prosodically marked sentences (as implemented by us), must result in local readings (in our task). The only data point that is relevant for our conclusions is that local readings arose even in the absence of stress on *some*. That is all we need, and all we can have. Since there is no point at which we make strong claims about the absence of local readings despite prosodic markedness, we do not discuss this point, also in the interest of length.

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).

As requested, we added references about the interaction of prosodic phrasing and modifier attachment. We also agree with Reviewer 1 that the two prosodic manipulations cannot be directly compared to each other. However, by choosing a prosodic realization that has been repeatedly shown to affect attachment decisions, we intended to “be on the safe side” when testing the sensitivity of the ITV task towards prosodic cues, even if these cues themselves differ between the constructions. Thus, we wanted to ensure that (potentially) absent prosodic effects for our target conditions cannot be attributed to a *general* lack of informativity of prosodic cues in the present task.

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. The 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 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. However, 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 each 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.

Thank you for pointing out this issue. The new version of the paper tries to fully acknowledge the reviewer’s well-taken points by (i) completely 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 suggestion made by Reviewer 2 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 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 focusing on the SMH. 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 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 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 that stood out as salient enough on conceptual grounds above other potential variants of grammaticalism in general or of Magri’s approach within it.

How would these interpretation strategies be implemented in Magri’s proposal? Under his approach – which is motivated by data, not *a priori* parsing principles – a complex sentence will have an EXH at each scope site. The literal meaning follows by treating alternatives as irrelevant, and hence by pruning them so that EXH is vacuous. So one strategy, call it the pruning strategy, is to prune alternatives from the domain of EXH. A second strategy, the strengthening strategy, is one in which you don’t prune alternatives at all. Thus, the only two form-meaning pairs predicted under Magri (2009, 2011), once a single auxiliary assumption is made about interpretation strategies, gives rise to precisely the two readings found in the authors’ data. What is needed, of course, is a statement about why the pruning strategy should be preferred to the strengthening strategy, and there are obvious suggestions worth pursuing, some of which the authors discuss (e.g., literal meanings require less computation, which might be preferred in a demanding task such as the one discussed in the authors’ paper). There is also the question of the differential availability of the local reading between *every* and *exactly-one*; this just means there’s work to do, but I don’t see how this is especially problematic for the grammatical theory. If anything, the data seem to support the approach, given the existence of a ready explanation that combines independently motivated assumptions about grammar (Magri, 2011) and independently motivated assumptions about interpretation strategies (Noveck and Posada, 2003).

We have picked up this intriguing suggestion in Section 6 and tried to strengthen the reviewer’s point even further. Even the observed preference order can be explained if we additionally assume that ES-sentences do not have the “global alternative” because that one does not asymmetrically entail the to-be-interpreted ES-sentence.

I may of course have miscalculated, and the authors are encouraged to double- check everything I’ve said above. The main point nevertheless stands: 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

We have incorporated this point by expanding on the interpretation of our data in Section 6.

As with grammaticalism, I feel that the authors are too quick to select from a set of certain choice points made available by traditionalism and take these choices as representative. For example, 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.

This seems to be a misunderstanding of what we had meant by ‘default’ in this context, and it is our fault that it arose. Thank you for pointing this out. We have reformulated the entire Section 3 with this in mind. The term ‘default’ is omitted entirely.

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.

We carefully checked our manuscript and could not find a place where we explicitly made this point. We decidedly do not want to suggest that, because we agree with the reviewer that it would be unjust. If you could point out to us where in the paper this unfortunate misunderstanding arose, we would be happy to improve the text accordingly.

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?

We stick to the term “traditionalism” and are more specific about what we mean with it in Section 3. There, we also address the reviewer’s worry expressed here that only stronger alternatives should matter. It is only the Neo-Gricean position that appeals to lexicalized scales for which what Chemla & Singh (2014, Compass) call *Stonger Alternatives* is relevant. Our more encompassing notion of ‘traditionalism’ is not strictly committed to that. It is true that dispensing with the *Stronger Alternatives* assumption gives rise to problems for traditionalism, but that does not mean that it cannot plausibly assume the mentioned alternative (8) to squeeze out a local reading for AS-sentences. This is also the stance adopted by, for example, Chemla & Spector in their experimental paper on the subject.

(1) p.9: ‘The grammatical approach, as described so far, is not yet a fully articulated theory.’

Theory of what? Theory of grammar, or theory of human behaviour? Would anyone say that a syntactic theory that posited PP-attachment ambiguities was not a ‘fully articulated theory’? Of course, these ambiguities need to be resolved somehow, and auxiliary assumptions about processing, memory, planning, etc. are invoked to say how. But this is commonplace, no? I somehow fail to see what the substance of this claim is. (The discussion in Chemla and Singh (2014a,b) might be relevant.)

This unfortunate locution was removed entirely, and the reviewer’s suggestion to stick to Chemla & Singh was taken up.

(2) A potential difference between EC/LC and literal/strengthened ambiguities

The authors found with the EC/LC ambiguities that there was a slight preference for delaying the response. This seems to be a rational strategy for a participant to follow – when there is an ambiguity it might make sense to wait for some time for relevant information to come in before deciding how to disambiguate, especially if there is little cost to doing so (here resolving the ambiguity merely requires an answer to the question: where should the PP be attached?). With scalar sentences, on the other hand, we might naturally expect a preference for literal readings when the task imposes high demands on memory, because the literal reading requires strictly less computation: no need to generate alternatives, no need to exhaustify, etc. A lazy participant might wish to avoid all that computation. I wonder if this consideration has any merit, and whether it bears on the interpretation of the results.

This is taken up in the discussion section. Thanks for bringing this up!

(3) Traditionalism

On p.6, the authors write that they use the word ‘traditionalism’ because it takes a ‘conservative stance’ with respect to Grice’s work. This totally ignores arguments by Fox (2007, 2014) that the most conservative understanding of the Maxim of Quantity only yields ignorance inferences, not SIs. It is only with a radical stipulation, grammatically stipulated formal alternatives, that a quantity maxim can be used to derive SIs. Furthermore, as Larry Horn has pointed out in many papers, the debate over the source of the apparent ambiguity is very old, much older than the modern era. In light of these considerations, calling the neo-Gricean approach ‘traditional’ does not entirely do justice to tradition.

Why not use labels that more accurately characterize the debate? Isn’t the issue one about the domain-specificity of the strengthening mechanism?

We do not feel strongly about the term ‘traditionalism’. We have kept it here, because it is not as loaded and misleading as other terms would be (such as ‘domain-general’). The ‘tradition’ that was started by Grice is enough of a historical tradition for our concerns. As pointed out above, this is not Neo-Griceanism as such, but something more in the spirit of Grice: a rationalization of speaker utterances.

But, of course, if the reviewer and/or the editors feel strongly about this issue, a different term can be found for certain.

I’m wondering if the authors might say a little bit more about how the incremental verification task suppresses typicality reasoning. Are the authors assuming that participants will not categorize certain configurations along the way as (a)typical? (Of course, I agree with the authors that ‘typicality’ is not an explanation but something in need of explanation; nevertheless, perhaps some comment might help further motivate the use of incremental verification.)

We address this issue in Section 6. Our main point is that for the aspired ‘typicality’-based explanation of in particular Geurts and van Tiel to apply, subjects would need to have expectations about likely typicality of pictures that they cannot fully see yet.

Not a particularly relevant question, but just wondering: did the authors by any chance collect RTs at critical regions? (might be informative)

We did collect timing information, but these were not always reliable enough. Importantly, given that there was no time pressure involved in the present task, the informativity of RTs is inherently restricted. As the paper is long and intricate already, we decided not to discuss RTs.

I fail to see the prior plausibility of the disambiguation criterion the authors state at the bottom of p.39.

We have removed this in favor of a discussion of Magri’s proposal.

p.39, last paragraph: The QUD criterion in Fox (2007) does make clear pre- dictions. The problem is not with the predictions, but with the unfortunate fact that QUDs are hard to pin down. If we could find a way to overcome this obstacle, the predictions of the theory are clear (e.g., one might try fixing the QUD and see what happens; without an explicitly provided QUD, listeners have to guess one, cf. also a similar point raised in connection with traditionalism in section 2.2).

We have tried to be clearer about our position on this in Section 3 and Section 6. The problem with the QUD based account is that it does not give non-trivial predictions for a neutral experiment that does not fix a QUD. In other words, the QUD approach does make testable predictions in the sense that there are experiments that can test it. But it does *not* make predictions about every experiment.

It can derive any relevant reading by the QUD whether that reading is true. But it does not predict which QUD is likely to be accommodated, at least not as far as we are aware of.

Typos and errors mentioned.

Fixed. Thanks!!

## Reviewer 3

[T]he methodology itself raises questions regarding the interpretation strategy participant may have adopted in order to complete the task at hand. What guided participants may have been independent of the actual salient and preferred readings and therefore may have biased them in a way that could have masked what the actual preferred readings were. More specifically: 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.

Thank you for addressing this issue. We have now expanded on alternative interpretations and potential confounds in a new dedicated discussion section (Section 6). The specific concern mentioned here, the “quick on the trigger” strategy is addressed in several ways. For one, the preference-related controls are supposed to control for such effects: this case shows that it is not generally so that participants exit at the first possible reading. Secondly, our data from the critical conditions also are not compatible with this idea. Finally, we present a short additional study on unembedded implicatures that also does not confirm this alternative interpretation but rather closely reflect the ratio of pragmatic vs. literal answers that are also found in plain picture-verification tasks.

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. The addition of such discussion will strengthen the article, sharpen the analysis of the data by ensuring that the authors understand what guided participants in the task, and will instill confidence in other scholars that the new paradigm is well understood and that response patterns that come out of using it are clear and replicable.

We have included a stand-alone discussion section to address the most pressing issues.

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.

This is an excellent discussion point which we address is Section 6. The upshot is that we cannot rule out that controls and critical conditions behave differently in terms of exiting biases in our sequence of pictures, but our experiment on unembedded implicatures undermines the idea that the contrast is very strong. And if it is only mild, our conclusions should be stated carefully enough so that they would still hold.

If the reviewer has a stronger processing difference in mind (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 would be to highlight these potential differences and to pave the way for subsequent processing-oriented studies that can address these worries.

The hypothesis that participants go with the most inclusive reading that is compatible with the pictorial information seems to account for the responses in the *exactly one. . . some*. In this condition, the first reading participants can make a decision about is the global one, which, given step 2 in Figure 5, is false. And so, the only reading the entails the other two readings is the first one out. Step 3 then rules out the literal reading. If the local reading is still available, then why didn’t participants wait to see if it is indeed true as more pictorial information is revealed? Here the account the authors provide, that the literal reading is the preferred one (followed, at least in this condition, by global and then the local), is convincing: Participants waited until the preferred reading is falsified (or, in all...some, confirmed) to make a decision.

Just to avoid misunderstanding: there is no most inclusive reading for ES-sentences. The local reading is logically independent of the other two.

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 are not totally sure whether we understand this alternative explanation correctly, and whether there would be a theoretical motivation for such a strategy. We believe that checking that *exactly n* is true necessarily requires to consider what property these exactly n objects are supposed to have. Essentially, such a strategy would be indistinguishable from what we address as ‘spill-over’ readings. Our hunch is that it is similar in thrust to what we suggest as a potential route for traditionalism to deal with the data from ES-sentences.

On a more general note, participants’ tendency to choose a construal that doesn’t rule out stronger readings is the opposite of what many analysis of conversational implicatures assume: that comprehenders would likely assume an informationally stronger, i.e. more restrictive, reading rather than a weaker one. The authors’ thoughts on these surprising findings are also welcome in the general discussion.

We strongly agree with this point and think that it in fact strengthens our case. This is exactly why we picked the Strongest Meaning Hypothesis as an ‘auxiliary assumption’ for grammaticalism. Unfortunately, we do not presently have anything substantial to add expect wild speculation about why the SMH would not be a good selection criterion. As the paper is already long and complex, we feel that this is better deferred.

Related to the concerns raised in the previous section, a note on the authors discussion of what the QUD is in their experiment. As they say, the QUD would be a very artificial concept here. 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.

This is an interesting proposal, but we do not see how to motivate it as a highly salient, theoretically motivated option at the outset of the experiment. There are many ways that grammaticalism can be made compatible with the data, once we know what the data is. We discuss one such way, proposed in more detail by another reviewer, and feel that it would not improve the paper if we were to discuss yet another option.

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.

Let us clarify that 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.

The modifications the authors seems incongruous. Could the authors suggest modifications for each view that would account for both the AS and ES accounts?Maybe there is an ambiguity of uniqueness and non-uniqueness reading for numerals, modified or otherwise. If such an ambiguity story works, the authors may want to see whether it stems from a unique/non-unique ambiguity of the quantity expression much/many that combines with the numeral, along the lines of Nouwen 2010’s S&P paper Two kinds of modified numerals. (See also Geurt’s 2006 paper *Take five*.)

We tried to make clearer now that we provide a uniform modification for each account so as to deal with AS-sentences and ES-sentences alike. In the revised version, there is a new uniform explanation for grammaticalism that builds on a reviewer’s comments that is similar in spirit, but even more thought-provoking than what we had proposed before. As for traditionalism, we still conclude that the only thing that traditionalism needs to ‘explain away’ is the strange behavior of ES-sentences.

As for the more general question whether the unexpected readings of ES sentences could principally be explained in such way that both traditionalists and grammaticalists might be satisfied, we believe that this is an intriguing issue. However, we fear that a solution to this question cannot be drawn based on the present findings.

1. Note that the vast majority of all studies on PP attachment focuses on *Minimal Attachment* ambiguities rather than *Late Closure* structures. In Minimal Attachment constructions, the attachment choice has to be made between a VP and an NP (such as in *Tap the frog with the flower*), whereas in *Late Closure* structures, the potential attachment sites are two noun phrases. [↑](#footnote-ref-1)