2023. *Proc Ling Soc Amer* 8(1). 5537. https://doi.org/10.3765/plsa.v8i1.5537.

Degree estimates as a measure of inference calculation

Eszter Ronai & Ming Xiang*

Abstract. Scalar inference (SI), e.g., utterances containing the quantifier *some* being enriched to mean *some but not all*, is a central topic in semantics-pragmatics. Of recent interest in the experimental literature is the phenomenon of scalar diversity: that different lexical scales exhibit variation is how likely they are to lead to SI. However, studies of scalar diversity have almost exclusively relied on a particular experimental task: the inference task. In this paper, we argue that the inference task suffers from a number of shortcomings: namely, that it biases by providing participants with the stronger alternative and that it obscures pragmatic inferences other than SI. Instead we offer as an alternative a degree estimate task to investigate utterances containing scalar terms. We use the degree estimate task to reassess previous inference task-based findings from the literature on how two manipulations (discourse context and *only*) affect the likelihood of inference calculation. Our results show that the two tasks produce results that differ from each other in subtle but important ways.

Keywords. scalar inference; scalar diversity; inference task; Question Under Discussion; focus semantics

1. Background.

1.1. SCALAR INFERENCE AND SCALAR DIVERSITY. Scalar inference (SI) is the phenomenon whereby sentences containing scalar terms are interpreted as conveying a strengthened, upper-bounded meaning. The best studied examples of SI are sentences involving the quantifier *some*, as in (1-a).

(1) a. Sue ate some of the cookies.
b. SI: Sue ate some, but not all, of the cookies.

One standard view on how SI arises is that comprehenders reason about informationally stronger unsaid alternatives, such as *Sue ate all of the cookies*. This sentence is an informationally stronger alternative to (1-a) because it asymmetrically entails it (Horn 1972). Since the speaker of (1-a) should have uttered the stronger alternative if she had been in a position to do so (Maxim of Quantity, Grice 1967), comprehenders can infer its negation (Maxim of Quality). Combining the negation of the stronger alternative (*Sue did not eat all of the cookies*) with the literal meaning of (1-a) (*Sue ate at least some of the cookies*) leads to the SI-enriched interpretation in (1-b).

While it has long been acknowledged that many other lexical items also form scales (i.a., Horn 1972; Hirschberg 1985), only relatively recently has attention turned to the experimental study of a wider range of scales —with the first large-scale investigation conducted by van Tiel et al. (2016), though see also Doran et al. (2012); Baker et al. (2009); Beltrama & Xiang (2013) for earlier work. Similarly to (1-a), an utterance of (2-a) can also trigger SI via the same reasoning process outlined above. Upon encountering (2-a), comprehenders reason about and derive the

* We would like to thank the audience at the 2023 Annual Meeting of the LSA, and especially Larry Horn, for helpful discussion. This material is based upon work supported by the National Science Foundation under Grant No. #BCS-2041312. All mistakes and shortcomings are our own. Authors: Eszter Ronai, Northwestern University (ronai@northwestern.edu) & Ming Xiang, The University of Chicago (mxiang@uchicago.edu).

© 2023 Author(s). Published by the LSA with permission of the author(s) under a CC BY license.

negation of the unsaid informationally stronger alternative *The movie is excellent*, leading to the SI-enriched meaning given in (2-b).

(2) a. The movie is good.
b. SI: The movie is good, but not excellent.

The influential finding in experimental studies of such different scales is that, although the reasoning process is identical, the likelihood of comprehenders deriving the SI is actually hugely variable. For instance, van Tiel et al. (2016) (Experiment 2) found that while almost 90% of participants calculated the *some but not all* SI, the rate of SI calculation for *good but not excellent* was less than 40%. In fact, the rate of calculation across the 43 different scales tested ranged from 4% to (almost) 100%. This robust variation has been termed *scalar diversity*. Work on scalar diversity has since concentrated on trying to identify factors, mostly related to properties of the different lexical scales, that can predict the likelihood of SI calculation from a given scale and explain the variation (van Tiel et al. 2016; Gotzner et al. 2018; Sun et al. 2018; Westera & Boleda 2020; Pankratz & van Tiel 2021; Ronai & Xiang 2021, 2022b).

1.2. THE INFERENCE TASK. Existing experimental work on scalar diversity has employed the so-called *inference task* to measure the likelihood of SI calculation. In this type of two-alternative forced choice task, participants are presented with sentences such as "Mary: *The movie is good*" and are asked the question "Would you conclude from this that Mary thinks the movie is not excellent?" —see Figure 1. Participants can then respond with "Yes" or "No". A "Yes" response indexes SI calculation, i.e., that the participant has computed the *good but not excellent* meaning of *good*. A "No" response indicates that the participant has not calculated the SI, and *good* is interpreted as *at least good*, which is then compatible with *excellent*.

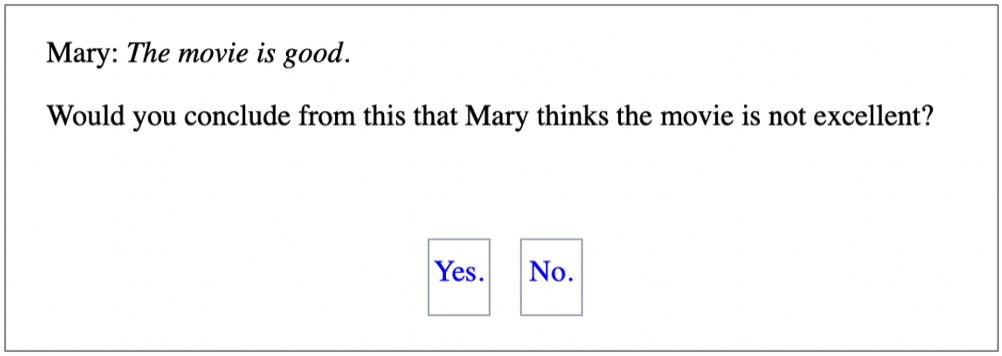

Figure 1. Inference task

As mentioned, the inference task has been widely used to study scalar diversity, with van Tiel et al. (2016) and all subsequent studies cited above relying on it. (While Sun et al.'s (2018) experiment asks for judgments on a 0-100 scale instead of a binary response, it is still an inference task with the same task question.) While existing research has undoubtedly uncovered interesting results using the inference task, we would like to argue that this method also has some shortcomings. First, the task question (*Would you conclude from this that Mary thinks the movie is not excellent?*) explicitly provides the stronger alternative (*excellent*), making it maximally salient. This might create a bias for participants to reason about that alternative, in turn biasing them toward calculating the SI. Second, there is a possibility that when a participant responds

with "Yes", other pragmatic inferences also affect their judgement. For instance, the relevant scalar terms may also undergo negative strengthening. Negative strengthening is the inference whereby comprehenders take *The movie is not excellent* to mean not only that the movie is less than excellent (the literal meaning), but that it is less than good, or in fact mediocre (Horn 1989). If participants calculate such an inference, then whether they respond with "Yes" vs. "No" no longer merely reflects whether they calculated SI from *The movie is good*, but whether they have negatively strengthened *...the movie is not excellent*.

There exist previous studies on SI (though not on scalar diversity) that have probed the appropriateness of experimental tasks and found differences among them. Geurts & Pouscoulous (2009) compared the inference task to the verification task. In the former, participants were provided with a statement like "Some of the Bs are in the box on the left." and had to answer a question such as "Would you infer from this that not all the Bs are in the box on the left?" with "Yes" vs. "No". In the latter, participants had to decide whether the same sentence ("Some of the B's are in the box on the left") correctly describes a picture where in fact all of the B's are in the box on the left. (Note that while, as before, in the inference task a response of "Yes" is what corresponds to SI calculation, in the verification task it is a response of "No".) The authors found that the inference task led to more robust calculation of the *some but not all* SI, at a rate of 62% (in Experiment 2), while the verification task led to SI at a rate of only 34%. Recently, Sun & Breheny (2022) compared two different versions of the task question for an inference task, one where the stronger alternative is embedded under negation vs. one where it is embedded under a possibility modal. In their experiments, participants were presented with an utterance like "Mary says: *Some of the questions are easy.*" and either had to respond to "Would you conclude from this that, according to Mary, not all of the questions are easy?" (negation) or to "Would you conclude that, it could be that Mary thinks, all of the questions are easy?" (modal). Results revealed significant differences between the different versions of the task question: for the <*some*, *all*> and <*possible*, *certain*> scales, the negation question resulted in more SIs, while for numerals, the modal question resulted in more SIs. (Though it must be noted that many have argued that numerals differ from standard cases of SI; see Koenig 1991; Breheny 2008; Solt & Waldon 2019 among many others.)

Existing work has also tested the effect of different numbers of response options on experimental outcomes (Katsos & Bishop 2011; Jasbi et al. 2019; Sikos et al. 2019). In particular, Jasbi et al. (2019) conducted sentence-picture verification studies, varying how many potential responses participants could choose from: two (wrong, right), three (wrong, neither, right), four (wrong, kinda wrong, kinda right, right); or five (wrong, kinda wrong, neither, kinda right, right). They found that the number of options had an effect on results, additionally raising the question of which response(s) should be taken to index SI calculation: a response of "wrong" or any response other than "right".

1.3. CONTRIBUTIONS OF THIS STUDY. In this paper, we employ a different experimental measure to test SI calculation in the context of the scalar diversity phenomenon. Our task measures which world states comprehenders come to have in mind, given an inference-triggering utterance such as *The movie is good*. Specifically, we collect degree estimates on the underlying degree scales, tapping into what degree of goodness comprehenders end up attributing to the movie, after encountering *The movie is good*, or the *The movie is only good*, etc. This provides a much more fine-grained measure than the binary inference task ("Yes" vs. "No"), and it also avoids the bias

of directly presenting participants with the stronger alternative. To serve as a reality check, in Experiment 1 we test utterances containing weaker scalar terms, stronger alternatives, and—in light of recent experimental findings about negative strengthening (Ruytenbeek et al. 2017; Gotzner et al. 2018)—negated stronger alternatives. In Experiment 2, we test the effect of the Question Under Discussion (QUD, Roberts 1996/2012) on inference calculation, as well as what happens when the tested sentences include the focus particle *only*. To briefly preview our findings, Experiment 2 finds effects that are subtly different from the results of previous experiments that tested the same two manipulations using an inference task (Ronai & Xiang 2022a). We interpret these differences in light of the two shortcomings of the inference task we raised above, namely that it creates a bias by presenting participants with the stronger alternative, and that it obscures the role of pragmatic inferences other than SI.

2. Experiment 1. In order to validate our methodology, we first used the degree estimate task to compare weaker scalar terms to their stronger alternatives, since we had a clear prediction that the former would lead to lower degrees than the latter. Additionally, given that the possibility of negative strengthening is a concern we raised for the inference task, and previous experimental work has been able to detect when participants calculate this inference (Ruytenbeek et al. 2017; Gotzner et al. 2018), Experiment 1 also tested negated stronger alternatives.

2.1. PARTICIPANTS AND TASK. 91 native speakers of American English participated in an online experiment, administered on the Ibex platform (Drummond 2007). Participants were recruited on Prolific and compensated $2. Native speaker status was established via a language background survey, where payment was not conditioned on participants' responses. Data from all 91 participants is reported below.

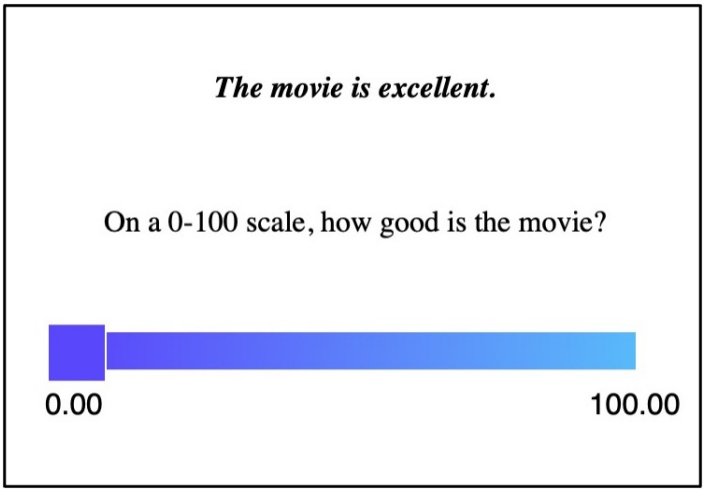

Figure 2. Example experimental trial from Experiment 1: strong scalar term condition

Experiment 1 used a degree estimate task. As mentioned, we tested the weaker scalar term (e.g., *good*), the stronger alternative (*excellent*), and the negated stronger alternative (*not excellent*). These three conditions were tested in a between-participants design (with 31 participants in the negated strong condition, and 30 participants each in the weak and strong conditions). Participants were presented with a speaker's utterance such as *The movie is good*, *The movie is excellent*, or *The movie is not excellent*. They were then asked the question "On a 0-100 scale, how

good is the movie?", and had to make a judgement by picking a point on a sliding scale. Figure 2 illustrates the task with an example from the strong scalar term condition. We aimed to create neutral task questions that would not bias participants toward either end of the scale. For adjectival lexical scales, for instance, questions relied on the weaker term wherever possible (e.g., *On a 0-100 scale, how old is the house?* for <*old, ancient*>) —see Section 5.2 in Ronai & Xiang (2022b) for additional details.

The experiment included 60 critical items, that is, 60 different lexical scales (adjectives, verbs, adverbs, quantifiers, and connectives). Section 2 in Ronai & Xiang (2022b) reports in more detail on the corpus work carried out to construct this scale set. 3 practice trials and 5 filler items were also included. The latter served as catch trials and used words in the sentence and task question that were each other's antonyms, e.g., *The table is clean* was paired with *On a 0-100 scale, how dirty is the table?*.

2.2. HYPOTHESES AND PREDICTIONS. Assuming that participants calculate SI (at least some of the time), we expect lower degree estimates, i.e., lower degrees of goodness attributed to the movie, given an utterance of *The move is good* (weak scalar condition) than an utterance of *The movie is excellent* (stronger alternative condition). If participants never calculate SIs like *good but not excellent*, then it is in principle possible that the weak scalar and stronger alternative conditions would not differ, since the literal, non-upper-bounded meaning of *good* is compatible with *excellent*[1].

The negated stronger alternative condition (*The movie is not excellent*) should receive lower degree estimates than the stronger alternative condition (*The movie is excellent)* based on the semantic contribution of negation. Moreover, if participants derive the negative strengthening inference, then the negated strong condition is predicted to result in degree estimates lower than even the weak scalar condition, since in that case, *The movie is not excellent* would end up meaning that the movie is less than good. Previous experimental work has shown that participants are indeed sensitive to negative strengthening. In Gotzner et al.'s (2018) study, participants saw sentences such as *He is not brilliant* and were asked whether they can conclude that he is not intelligent. The authors found evidence for negative strengthening, i.e., "Yes" responses (see also Ruytenbeek et al. 2017). If our degree estimate task is similarly able to identify negative strengthening, then the negated strong condition should lead to the lowest degree estimates in Experiment 1.

2.3. RESULTS AND DISCUSSION. Figure 3 shows the results of Experiment 1 as violin plots. For the statistical analysis, we fit a linear mixed effects regression model using the lme4 package in R (Bates et al. 2015). The model predicted Response (0-100) by Condition (weak vs. strong vs. not strong). The fixed effects predictor Condition was treatment-coded, with weak as the reference level. Random intercepts were included for participants and items. Responses to strong terms were found to be significantly higher than to weak terms (Estimate=22.68, Std. Error: 2.68, t=8.48, $p<0.001$). Responses to negated strong terms were found to be significantly lower than to weak terms (Estimate=-33.59, Std. Error: 2.65, t=-12.65, $p<0.001$).

[1] In reality things are more complicated than this, since both *good* and *excellent* denote intervals, yet in the degree estimate task we ask participants to pick a single point. If SI from *good* has not been calculated, then the intervals denoted by *good* and *excellent* are overlapping, so participants may indeed pick points for *good* that are as high as *excellent*. However, if we assume that for an interval, participants pick the middle point, then *good* would result in lower degree estimates than *excellent* even if SI was not calculated, since the interval does start lower.

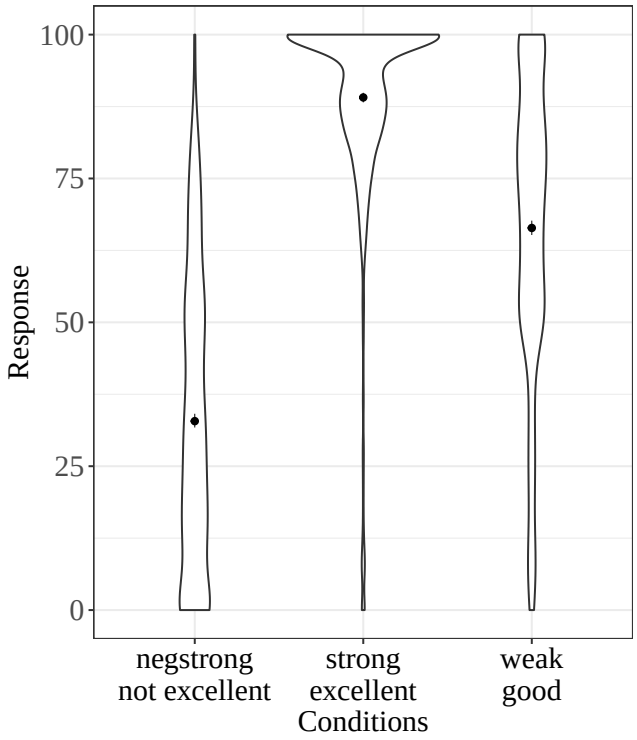

Figure 3. Experiment 1 results. Dots represent means and error bars 95% confidence intervals.

Our first finding, then, is that averaged over all critical items, stronger alternatives received higher ratings than the weaker terms. In other words, a sentence such as *The movie is excellent* led hearers to attribute a higher degree of goodness to the movie than *The movie is good*. This result serves as a reality check and as confirmation that participants were performing the task adequately. Additionally, this can be taken as evidence that participants calculated SIs like *not excellent* from *The movie is good* —though see footnote 1 for a brief discussion of complications for this interpretation. We return to the question of what degree estimate findings would correspond to SI calculation in Section 4.

Secondly, we found that sentences such as *The movie is not excellent* received, on average, lower ratings on a 0-100 goodness scale than sentences such as *The movie is good*. That is, sentences such as *The movie is not excellent* led participants to believe that the movie is less than good. This can be interpreted as negative strengthening (Horn 1989), confirming that our experimental paradigm is able to detect such pragmatic inferences. Negative strengthening will also be relevant in our interpretation of some of the Experiment 2 findings (Section 3.3).

3. Experiment 2. Having used Experiment 1 as a basic validation of the degree estimate task, in Experiment 2 we use this task to reassess previous findings from experimental work that used the inference task. Specifically, we look at how the likelihood of inference calculation changes when sentences like *The movie is good* appear in a discourse context (more specifically, answering a QUD, operationalized as an explicit question), or when they include the focus particle *only*.

3.1. PARTICIPANTS AND TASK. 97 native speakers of American English participated in an experiment on the Ibex platform, for either $2 (*only* experiment) or $2.25 (QUD experiment) com-

pensation. Participant recruitment and screening was identical to Experiment 1. A total of 5 participants were excluded from analysis for failing attention checks (fillers). For the *only* experiment, data from 32 participants is reported; for the QUD experiment, data from 60 participants is reported.

In Experiment 2, we modified sentences from the weak scalar condition of Experiment 1 in the following ways. First, we placed sentences in a dialogue context, where inference-triggering sentences were preceded by a polar question that contained either the stronger alternative ((3), strong QUD condition) or the weaker scalar term itself ((4), weak QUD condition). The inference-triggering sentences were modified to ensure dialogue coherence, e.g., in Mary's utterance *The movie...* was changed to *It...*; otherwise, they were identical to Experiment 1. The QUD manipulation was administered within-participants.

(3) Sue: Is the movie excellent?
Mary: It is good.

(4) Sue: Is the movie good?
Mary: It is good.

The third condition in Experiment 2 modified the Experiment 1 weak scalar sentences such that they now included the focus particle *only* ((5), *only* condition). The *only* condition was tested as a between-participants manipulation.

(5) The movie is only good.

Experiment 2 was otherwise identical to Experiment 1 in its items (critical, practice, fillers), instructions, task questions (On a 0-100 scale, how good is the movie?), and procedure. Figure 4 shows an example trial from the strong QUD condition.

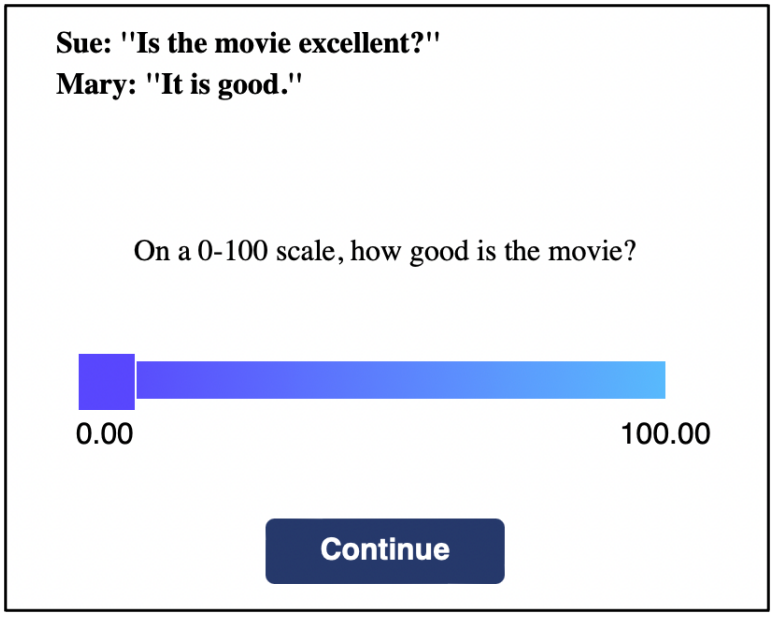

Figure 4. Example experimental trial from Experiment 2: strong QUD condition

3.2. HYPOTHESES AND PREDICTIONS. For our predictions, let us turn to findings from Ronai & Xiang (2022a), who used the inference task to test conditions identical to the current Experiment 2 (sentences like (3)-(5)). First, Ronai & Xiang (2022a) found that the weak QUD condition (4) led to the same SI rate as contextless sentences (*The movie is good*) —therefore, we expect to be able to use this condition as a baseline as well. Second, the authors report that SI rates were significantly higher across the board in a supportive discourse context (strong QUD condition): e.g., the *good but not excellent* SI was more likely to arise in (3) than in either (4) or its contextless version (2-a). Third, the presence of the focus particle *only* (5) resulted in inference rates even higher than the supportive context of the strong QUD condition. As mentioned, these findings come from experiments that used the inference task, where an increase in "Yes" responses (from (4) to (3) to (5)) was taken to index increased inference calculation.

These findings can be offered the following explanation. A biasing question encourages SI calculation, which leads to higher observed rates —see i.a., the Question-Answer Congruence proposal of Hulsey et al. (2004), as well as experimental findings from Degen (2013); Zondervan et al. (2008). However, this is merely a pragmatic effect. The focus particle *only*, on the other hand, encodes the exclusion of alternatives semantically, in the grammar (Rooth 1985, 1992). Since the calculation of an upper-bounded meaning in this case is no longer a cancellable pragmatic inference, it is not surprising that inference rates would be (even) higher in this condition.

Overall, if the degree estimate task were to replicate findings from the inference task, we would expect to find most robust inference calculation (leading to the lowest degree estimates) in the *only* condition, followed by the strong QUD condition, with the baseline weak QUD condition leading to the least inference calculation, and therefore the highest degree estimates.

3.3. RESULTS AND DISCUSSION. Figure 5 shows the results of Experiment 2 as violin plots. Similarly to Experiment 1, for the statistical analysis, we fit linear mixed effects regression models using lme4. The model predicted Response (0-100) by Condition and included random intercepts for participants and items. The Condition predictor was treatment-coded, with the baseline weak QUD condition serving as the reference level. As compared to weak QUD, the analysis found significantly lower degree estimates for both the *only* (Estimate=-5.33, Std. Error: 2.6, t=-2.05, $p<0.05$) and strong QUD (Estimate=-12.2, Std. Error: 0.64, t=-19, $p<0.001$) conditions. For an additional pair comparison, we also fit a model where the *only* condition served as the reference level (but which was otherwise identical). This revealed that the strong QUD condition led to significantly lower degree estimates than the *only* condition (Estimate=-6.87, Std. Error: 2.6, t=-2.65, $p<0.01$). (Though significant, we note that the differences between conditions are smaller than in Experiment 1.)

To summarize, we found that the degree of goodness attributed to the movie was highest in the baseline context of *Is the movie good?*. It was lower given the sentence *The movie is only good*, and lowest given the dialogue context of *Is the movie excellent?*. This is a reversal of the findings obtained by Ronai & Xiang (2022a). Concretely, using the inference task, more *not excellent* inferences were found with *only* than with the strong QUD. Yet using the degree estimate task, a lower degree of goodness was found with the strong QUD than with *only* —where lower degrees correspond to more robust calculation of the *not excellent*-type inferences.

We offer two potential explanations for this reversal of findings. First, though *only* encodes the exclusion of alternatives semantically, it does not specify what those alternatives are. That is, *The movie is only good* can mean that the movie is not excellent, but it can also mean the exclu-

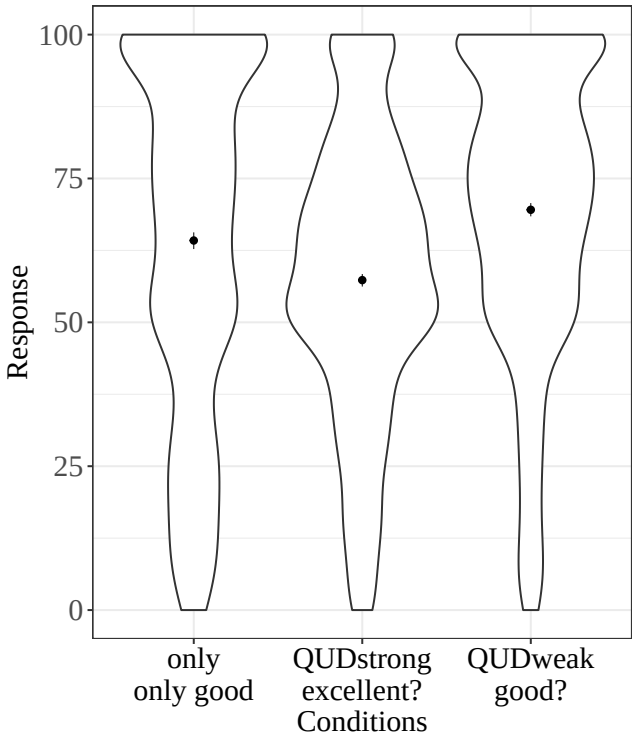

Figure 5. Experiment 2 results. Dots represent means and error bars 95% confidence intervals..

sion of non-scalar alternatives, e.g., that the movie is not funny. It is possible that participants in our Experiment 2 interpreted *only* as excluding alternatives that are not along the dimension of the degree scale (e.g., *funny*). This resulted in the degree estimate of goodness remaining higher. In the inference task, on the other hand, the task question ("Would you conclude from this that Mary thinks the movie is not excellent?") specifies the stronger scalar alternative *excellent* and forces participants to reason about that alternative. This may have the result of inflating the rates of calculating the *good but not excellent* meaning and the corresponding "Yes" responses.

Second, it is also possible that negative strengthening lowers the degree estimates for dialogues such as (6) (repeated from (3)).

(6) Sue: Is the movie excellent?
Mary: It is good.

In particular, it is possible to interpret such a dialogue as Mary intending to give a negative answer, but deciding not to overtly say "No" out of politeness. By saying *good*, however, Mary potentially actually intends to communicate *not excellent*. Mary's answer, then, can undergo negative strengthening via this indirect route, and it ultimately ends up meaning *less than good*. This results in a lower degree estimate. In the inference task, in contrast, no matter whether a participant thinks that Mary's answer in (6) means *good but not excellent* (the SI-enriched meaning), or that it in fact means *less than good* (via indirect negative strengthening), they will respond with "Yes".

One prediction of this idea to be explored in future work is that positive and negative polarity scales should diverge. Positive (adjectival) scales show a stronger negative strengthening

effect than negative ones —see i.a., Kritka (2007) for a theoretical proposal and Ruytenbeek et al. (2017) for supporting experimental findings. Intuitively this makes sense if negative strengthening is related to politeness considerations: saying a negated positive adjective (*not excellent*) instead of its antonym (*mediocre*) allows the speaker to save face, but this is not the case if polarity is reversed (Horn 1989; Brown & Levinson 1987). If our strong QUD results in the degree estimate task ultimately derive, in part, from politeness considerations and negative strengthening, then we should find differences across items according to polarity. Unfortunately, the current scale set is not best suited for such an analysis, since only a relatively small proportion of scales have negative polarity. Additionally, it is not obvious how non-adjectival scales could be classified as positive vs. negative; though see van Tiel & Pankratz (2021); Ruytenbeek et al. (2017) (and references therein) for how this provides a challenge even for adjectival scales.

In sum, the above two potential reasons might explain why in Experiment 2, we find lower degree estimates for (3) than for (5), which is counter to what has been found in previous work using the inference task.

4. Conclusion and open questions. An inference task is often used to test SI calculation; it is especially pervasive in investigations of scalar diversity. However, such a task is biasing, as it provides participants with a particular stronger scalar alternative, and it obscures what other non-SI inferences are factored into participants' "Yes" vs. "No" response. In this paper, we instead used a degree estimate task to test the role of a supportive context and *only* in modulating inference calculation. Our results were not entirely in line with previous work that used the inference task (Ronai & Xiang 2022a). Concretely, inference task findings had revealed that supportive contexts lead to an increase in SI calculation rates. But the likelihood of calculating inferences such as *good but not excellent* was in fact highest when sentences included the focus particle *only*. Using the degree estimate task, we instead found that inference calculation is most robust (that is, degree estimates are lowest) with a supportive discourse, and the results obtained with *only* fall in the middle between strong QUDs and the baseline condition. This highlights the value of using a more fine-grained, rather than binary, measure of inference calculation.

One important open question that remains is what corresponds to SI calculation in the degree estimate task. While in the inference task, it is clear that of the two response options, "Yes" indexes SI calculation, in the degree estimate task it is less obvious how we could tell whether a participant has calculated SI. As briefly discussed in Section 2.3, the finding in Experiment 1 that *good* elicited lower degree estimates than *excellent* could be taken as suggestive evidence that participants have calculated an upper-bound. However, the fact that both terms denote intervals complicates this interpretation. Therefore we cannot draw firm conclusions about whether SI calculation has happened from the degree estimate data of sentences like *The movie is good*. Another alternative would be to collect degree estimates on sentences like *The movie is good but not excellent* —in this case, we can of course be sure that the meaning participants are reasoning with is the upper-bounded *good but not excellent*. But in this case, the *not excellent* meaning is part of the asserted content, and therefore has a very different status from an SI. Therefore, neither of these two options represents an unproblematic candidate for tapping into SI calculation directly via the degree estimate task.

Nonetheless, given the inference task's shortcomings that we have discussed, and its virtual monopoly in the study of scalar diversity, we would still like to argue that degree estimates represent an interesting new way of looking at SI and scalar diversity.

References

Baker, Rachel, Ryan Doran, Yaron McNabb, Meredith Larson & Gregory Ward. 2009. On the non-unified nature of scalar implicature: An empirical investigation. *International Review of Pragmatics* 1(2). 211–248. https://doi.org/10.1163/187730909x12538045489854.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. https://doi.org/10.18637/jss.v067.i01.

Beltrama, Andrea & Ming Xiang. 2013. Is 'good' better than 'excellent'? An experimental investigation on scalar implicatures and gradable adjectives. In Emmanuel Chemla, Vincent Homer & Grégoire Winterstein (eds.), *Proceedings of Sinn und Bedeutung 17*, 81–98. Konstanz: University of Konstanz.

Breheny, Richard. 2008. A new look at the semantics and pragmatics of numerically quantified noun phrases. *Journal of Semantics* 25(2). 93–139. https://doi.org/10.1093/jos/ffm016.

Brown, Penelope & Stephen C Levinson. 1987. *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.

Degen, Judith. 2013. *Alternatives in pragmatic reasoning*. Rochester: University of Rochester dissertation.

Doran, Ryan, Gregory Ward, Meredith Larson, Yaron McNabb & Rachel E. Baker. 2012. A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language* 88(1). 124–154. https://doi.org/10.1353/lan.2012.0008.

Drummond, Alex. 2007. Ibex Farm. http://spellout.net/ibexfarm.

Geurts, Bart & Nausicaa Pouscoulous. 2009. Embedded implicatures?!? *Semantics and Pragmatics* 2(4). 1–34. https://doi.org/10.3765/sp.2.4.

Gotzner, Nicole, Stephanie Solt & Anton Benz. 2018. Scalar diversity, negative strengthening, and adjectival semantics. *Frontiers in Psychology* 9. 1659. https://doi.org/10.3389/fpsyg.2018.01659.

Grice, Herbert Paul. 1967. Logic and Conversation. In Paul Grice (ed.), *Studies in the way of words*, 41–58. Cambridge, MA: Harvard University Press.

Hirschberg, Julia Bell. 1985. *A theory of scalar implicature*. Philadelphia: University of Pennsylvania dissertation.

Horn, Laurence R. 1972. *On the semantic properties of logical operators in English*. Los Angeles: UCLA dissertation.

Horn, Lawrence R. 1989. *A natural history of negation*. Chicago: University of Chicago Press.

Hulsey, Sarah, Valentine Hacquard, Danny Fox & Andrea Gualmini. 2004. The question-answer requirement and scope assignment. In Aniko Csirmaz, Andrea Gualmini & Andrew Nevins (eds.), *MIT Working Papers in Linguistics*, 71–90. Cambridge, MA: MITWPL.

Jasbi, Masoud, Brandon Waldon & Judith Degen. 2019. Linking hypothesis and number of response options modulate inferred scalar implicature rate. *Frontiers in Psychology* 10. 189. https://doi.org/10.3389/fpsyg.2019.00189.

Katsos, Napoleon & Dorothy V.M. Bishop. 2011. Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition* 120(1). 67–81. https://doi.org/10.1016/j.cognition.2011.02.015.

Koenig, Jean-Pierre. 1991. Scalar predicates and negation: Punctual semantics and interval interpretations. *Chicago Linguistic Society (CLS)* 27. 140–155.

Kritka, Manfred. 2007. Negated antonyms: Creating and filling the gap. In Uli Sauerland & Penka Stateva (eds.), *Presupposition and implicature in compositional semantics*, 163–177. Houndmills: Palgrave Macmillan.

Pankratz, Elizabeth & Bob van Tiel. 2021. The role of relevance for scalar diversity: a usage-based approach. *Language and Cognition* 13(4). 562–594. https://doi.org/10.1017/langcog.2021.13.
Roberts, Craige. 1996/2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics* 5(6). 1–69. https://doi.org/10.3765/sp.5.6.
Ronai, Eszter & Ming Xiang. 2021. Exploring the connection between Question Under Discussion and scalar diversity. *Proceedings of the Linguistic Society of America (PLSA)* 6(1). 649–662. https://doi.org/10.3765/plsa.v6i1.5001.
Ronai, Eszter & Ming Xiang. 2022a. Quantifying semantic and pragmatic effects on scalar diversity. *Proceedings of the Linguistic Society of America (PLSA)* 7(1). 5216. https://doi.org/10.3765/plsa.v7i1.5216.
Ronai, Eszter & Ming Xiang. 2022b. Three factors in explaining scalar diversity. In Daniel Gutzmann & Sophie Repp (eds.), *Proceedings of Sinn und Bedeutung* (Vol. 26), 716–733. Konstanz: University of Konstanz.
Rooth, Mats. 1985. *Association with focus*. Amherst: University of Massachusetts dissertation.
Rooth, Mats. 1992. A theory of focus interpretation. *Natural Language Semantics* 1(1). 75–116. https://doi.org/10.1007/bf02342617.
Ruytenbeek, Nicolas, Steven Verheyen & Benjamin Spector. 2017. Asymmetric inference towards the antonym: Experiments into the polarity and morphology of negated adjectives. *Glossa: A Journal of General Linguistics* 2(1). 92. https://doi.org/10.5334/gjgl.151.
Sikos, Les, Minjae Kim & Daniel J. Grodner. 2019. Social context modulates tolerance for pragmatic violations in binary but not graded judgments. *Frontiers in Psychology* 10. 510. https://doi.org/10.3389/fpsyg.2019.00510.
Solt, Stephanie & Brandon Waldon. 2019. Numerals under negation: Empirical findings. *Glossa: A Journal of General Linguistics* 4(1). 113. https://doi.org/10.5334/gjgl.736.
Sun, Chao & Richard Breheny. 2022. The role of alternatives in the interpretation of scalars and numbers: Insights from the inference task. *Semantics and Pragmatics* 15(8). https://doi.org/10.3765/sp.15.8.
Sun, Chao, Ye Tian & Richard Breheny. 2018. A link between local enrichment and scalar diversity. *Frontiers in Psychology* 9. 2092. https://doi.org/10.3389/fpsyg.2018.02092.
van Tiel, Bob & Elizabeth Pankratz. 2021. Adjectival polarity and the processing of scalar inferences. *Glossa: A Journal of General Linguistics* 6(1). 32. https://doi.org/10.5334/gjgl.1457.
van Tiel, Bob, Emiel Van Miltenburg, Natalia Zevakhina & Bart Geurts. 2016. Scalar diversity. *Journal of Semantics* 33(1). 137–175. https://doi.org/10.1093/jos/ffu017.
Westera, Matthijs & Gemma Boleda. 2020. A closer look at scalar diversity using contextualized semantic similarity. *Proceedings of Sinn und Bedeutung* 24(2). 439–454. https://doi.org/10.18148/sub/2020.v24i2.908.
Zondervan, Arjen, Luisa Meroni & Andrea Gualmini. 2008. Experiments on the role of the question under discussion for ambiguity resolution and implicature computation in adults. In Tova Friedman & Satoshi Ito (eds.), *Proceedings of Semantics and Linguistic Theory (SALT) 18*, 765–777. Washington, DC: Linguistic Society of America. https://doi.org/10.3765/salt.v18i0.2486.