Language and Cognition (2025), 17, e17, 1–26
doi:10.1017/langcog.2024.55

ARTICLE

Scalar inference calculation through the lens of degree estimates

Eszter Ronai[1] and Ming Xiang[2]

[1]Department of Linguistics, Northwestern University, Evanston, IL, USA and [2]Department of Linguistics, The University of Chicago, Chicago, IL, USA
Corresponding author: Eszter Ronai; Email: ronai@northwestern.edu

(Received 20 June 2024; Revised 02 September 2024; Accepted 04 September 2024)

Abstract
Scalar inference (SI), e.g., utterances containing *some* being enriched to mean *some but not all*, is a central topic in semantics and pragmatics. Of recent interest in the experimental literature is scalar diversity: different lexical scales differ in their likelihood of leading to SI. Studies of scalar diversity have almost exclusively relied on the so-called inference task. In this article, we highlight two shortcomings of the inference task: it biases participants by providing them with the stronger alternative, and it obscures pragmatic inferences other than SI. We offer as an alternative a degree estimate task to investigate utterances containing scalar terms. We validate the degree estimate task, i.a., by successfully replicating a previous finding about scalar diversity: that the distinctness of scalar terms (*some* versus *all*) is a significant predictor of it. We then use degree estimates to reassess previous inference task-based findings. Our results show that biasing discourse contexts lead to lower degree estimates (i.e., more strengthened meanings) than a manipulation with *only*, which contrasts with prior literature's findings. The article concludes that the inference and degree estimate tasks both have advantages: the former offers a straightforward definition of SI calculation, while the latter avoids explicitly mentioning a negated stronger alternative.

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

1. Introduction

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 of how SI arises is the (neo-)Gricean account according to which 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., Hirschberg 1985; Horn 1972), 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 Baker et al. 2009, Beltrama and Xiang 2013, and Doran et al. 2012 for earlier work). Like (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 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.

An influential finding in experimental studies of such different scales is that, although the reasoning process is identical across scales, 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.*

A prominent research question regarding scalar diversity is: what properties of scales predict the likelihood of SI calculation? Existing work has identified a number of such properties. First, van Tiel et al. (2016) found that the distinctness of scalar terms, as operationalized by semantic distance and boundedness, successfully predicts scalar diversity (see also Westera and Boleda 2020). The authors also hypothesized that the availability of the stronger alternative would be a predictor; while this was not borne out in their own data, later work has provided some support for this hypothesis (Hu et al., 2022). Investigating adjectival scales, Gotzner et al. (2018) found that certain semantic properties of adjectives, such as polarity and extremeness, are also relevant for scalar diversity (see also Beltrama and Xiang 2013). Looking at the interaction of SI calculation and other inference types, Sun et al. (2018) showed that scalar diversity was positively correlated with local enrichability, i.e., with scales' propensity to lead to upper-bounded meaning in an embedded context, while Gotzner et al. (2018) showed a negative correlation with negative strengthening – we return to this latter inference type in more detail in our own experiments. Lastly, Pankratz and van Tiel (2021) related variation in SI rates to a usage-based notion of relevance, while Ronai and Xiang (2021) investigated the role of the Question Under Discussion in explaining it.

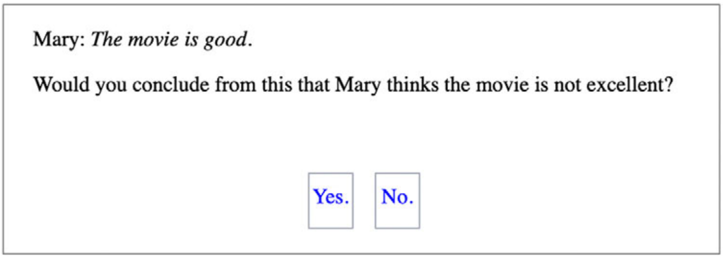
https://doi.org/10.1017/langcog.2024.55 Published online by Cambridge University Press

1.2. The inference task

While the majority of existing scalar diversity literature has focused on finding factors that can explain it, most important for our article is the experimental task that has tended to dominate in these studies. Prior experimental work on scalar diversity has employed the 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 for an example. Participants can then respond with 'Yes' or 'No'. A 'Yes' response is taken to index SI calculation, i.e., that the participant has computed the *good but not excellent* meaning of *good*. A 'No' response is taken to indicate that the participant has not calculated the SI, and *good* is interpreted as *at least good*, which is compatible with *excellent*.

As mentioned, the inference task has been widely used to study scalar diversity. It was first used by van Tiel et al. (2016) in their investigation of 43 different scales. Subsequent studies that tested similarly large sets of lexical scales and aimed to find explanations for scalar diversity either also used the inference task (Gotzner et al., 2018; Pankratz and van Tiel, 2021; Ronai and Xiang, 2021, 2024) or modeled data from studies that did use that task (Hu et al., 2022; Westera and Boleda, 2020). Note that there are slight variations in the task question that we gloss over here, such as the difference between *Would you conclude from this that Mary thinks the movie is not excellent?* versus *Would you conclude from this that, according to Mary, the movie is not excellent?* Lastly, Sun et al.'s (2018) experiment asked for judgments on a 0–100 scale instead of a binary response, but it was still an inference task with the same task question. As will be discussed in more detail below, two key properties of the inference task can be seen as shortcomings: first, it directly provides participants with the stronger scalar alternative it is testing (*excellent* in Figure 1), and second, it places that alternative under negation, which might lead to the intrusion of pragmatic inferences other than SI.

Before going into detail about why these properties represent potential shortcomings, we first note that there exist close variants of the inference task that have been used to test SI calculation across multiple lexical scales. These tasks still share at least one of the two key properties mentioned above. Cummins and Rohde (2015) used task questions of the form *How likely is it that the view is not gorgeous?* (given an utterance of *The view from the hotel room is pretty*) and asked for likelihood ratings on a 1–7 scale. Just like the inference task, this task presents participants with the negated stronger alternative (here, *not gorgeous*). De Marneffe and Tonhauser (2019) had



Mary: *The movie is good*.

Would you conclude from this that Mary thinks the movie is not excellent?

Figure 1. Inference task.

https://doi.org/10.1017/langcog.2024.55 Published online by Cambridge University Press

participants answer questions such as *Does Julie mean that her hike was exhausting?* (given dialogues like Mike: *Was your hike exhausting?* Julie: *It was strenuous*) on a 7-point scale that ranged from 'Definitely not' to 'Definitely yes'. Here, though the stronger alternative is not negated, it is still explicitly provided to the experimental participants via the task question. Most recently, Sun et al. (2023) have adopted Degen's (2015) paraphrase task to investigate whether testing a large sample of corpus-based stimuli leads to a reduction in the scalar diversity effect. This task asks participants to rate how similar an inference-triggering sentence (*Gaining full knee extension can be difficult after surgery*) is to a version with the SI inserted (*Gaining full knee extension can be difficult, but not impossible, after surgery*); this task also directly presents participants with the negated stronger alternative (*not impossible*).

Having established that the inference task – and close variants of it, which share at least some of its key properties – is the dominant method in studies of SI across different lexical scales, let us now expand on the advantages and disadvantages of this task. There are a number of key positive features of the inference task: it provides a straightforward operationalization of SI calculation ('Yes' responses), it easily generalizes across scales, and in its typical formulation, it asks participants to reason about speaker beliefs (e.g., *Mary thinks that* or *according to Mary*). And existing research using the inference task has undoubtedly uncovered interesting results. However, we would like to argue that this method also has some weaknesses. 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.[1] This might create a bias for participants to reason about that alternative, in turn biasing them toward calculating the SI. Indeed, Geurts and Pouscoulous (2009) – a study we review in more detail below – have argued that the inference task leads to elevated SI rates compared to other tasks, and provided experimental evidence supporting this. Second, there is a possibility that other pragmatic inferences also affect participants' interpretations of Mary's utterance and the task question, but the binary forced choice task is not able to pick up on this. For instance, the relevant scalar terms may also undergo negative strengthening. If the task question is interpreted with negative strengthening, then participants would 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 (a possibility briefly acknowledged by van Tiel et al. 2016, p. 149, see also Benz et al. 2018; Gotzner et al. 2018). If such an inference is indeed calculated, then whether participants respond with 'Yes' vs. 'No' no longer merely reflects whether they calculated the SI from *The movie is good*, but whether they have (also) negatively strengthened *not excellent* in the inference task.

[1] It must be acknowledged that work preceding van Tiel et al. (2016) tended to paraphrase, rather then directly mention, (non-negated) stronger alternatives; Baker et al. (2009) and Doran et al. (2012) used, for example, *the entire birthday cake* to correspond to *all* (of the cake) on the <*some*, *most*, *all*> scale, *the Lord of the Rings trilogy* to correspond to *three* (books) on the <*one*, *two*, *three*> scale, or *voted World's Most Beautiful Woman* to correspond to *gorgeous* on the <*average-looking*, *pretty*, *gorgeous*> scale. This method is also employed by Simons and Warren (2018), who tested the role of context on scalar diversity (albeit looking at only 9 scales) and paraphrased stronger alternatives using e.g., *100%* as stronger than *some*, *around 85–95 degrees* as stronger than *warm*, or *between 32 °F and 42 °F* as stronger than *cool.* Overall, however, the inference task (including its close variants) has become and remains the predominant way to test SI calculation in investigations of the scalar diversity phenomenon.\

https://doi.org/10.1017/langcog.2024.55 Published online by Cambridge University Press

There exist previous studies on SI that have shown that different experimental tasks can indeed yield different results, which is what the present article aims to do specifically in the context of scalar diversity. Geurts and Pouscoulous (2009) compared the inference task to the verification task. In the former, participants were provided with a statement like ‘Some of the B’s are in the box on the left.’ and had to answer a question such as ‘Would you infer from this that not all the B’s are in the box on the left?’ with ‘Yes’ versus ‘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 (in their Experiment 2) that the inference task led to more robust calculation of the *some but not all* SI at a rate of 62%, while the verification task led to SI at a rate of only 34%. Recently, Sun and Breheny (2022) compared two different versions of the task question for an inference task, one where the stronger alternative is embedded under negation versus 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 Breheny 2008; Koenig 1991; Solt and Waldon 2019 among many others.)

Other existing work has tested the effect of different numbers of response options on experimental outcomes (Jasbi et al., 2019; Katsos and Bishop, 2011; 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

We have argued that the widely used inference task (see Figure 1) faces some methodological concerns, relating both to the task question and the dependent measure. In particular, including the stronger alternative in the task question makes it more salient, potentially inflating rates of SI calculation, and only allowing participants to answer ‘Yes’ or ‘No’ may not perfectly index SI calculation rates, as answers may be influenced by other pragmatic inferences like negative strengthening. In this article, we introduce a novel method, which we call the *degree estimate* task, and we explore the utility of this task in studying the meaning of utterances containing scalar terms. Our degree estimate task assesses the properties of the world state(s) comprehenders come to have in mind, given an inference-triggering utterance such as *The movie is good.* Specifically, we collect degree estimates on the

https://doi.org/10.1017/langcog.2024.55 Published online by Cambridge University Press

underlying degree scales, tapping into what degree of goodness comprehenders attribute to the movie after encountering *The movie is good*, *The movie is excellent*, or the *The movie is only good*, and so forth. In this task, participants have to locate different scalar terms on a 0–100 scale. For example, to elicit degree estimates associated with *excellent*, we provide participants with the sentence *The movie is excellent* as well as a sliding scale with endpoints labeled 0 and 100, and ask them to respond to the prompt 'On a 0–100 scale, how good is the movie?'.[2] Importantly, the degree estimate task provides a more fine-grained measure than the binary inference task ('Yes' vs. 'No'), and it also avoids the bias of directly presenting participants with the (negated) stronger alternative.

To serve as a reality check, in Experiment 1 (Section 2) we test utterances containing weaker scalar terms and stronger alternatives. Additionally, in light of recent experimental findings about negative strengthening (Benz et al., 2018; Gotzner et al., 2018; Ruytenbeek et al., 2017) and the concern discussed above that the inference task may obscure this type of inference, Experiment 1 also tests negated stronger alternatives. We find that all three conditions result in reliably different degree estimates: a weaker scalar like *good* is lower on the scale of goodness than the stronger term *excellent*, while the negated stronger term *not excellent* is the lowest. We take these results as indicating that participants are able to interpret the degree estimate task and perform adequately, and as showing that the task is able to detect pragmatic inferences such as negative strengthening. Following this, we look at the by-scale variation in Experiment 1 (Section 3). First, we use degree estimates to successfully replicate van Tiel et al.'s (2016) finding that the more distinct two scalar terms (e.g., *good* and *excellent*) are from one another, the more robust the corresponding SI arising from the weaker one (*good but not excellent*). This serves as further validation of the degree estimate task. Second, we find evidence that the difference between the degree estimates of the negated strong term (*not excellent*), as compared to the degree estimates of the weak term (*good*), can predict the relative rates of 'Yes' responses in the inference task. This suggests that obscuring the effects of negative strengthening is indeed a shortcoming of that task.

Having established the above basic patterns that validate the use of degree estimates, in Experiment 2 (Section 4) we use this task to reevaluate previously obtained findings from inference task experiments (Ronai and Xiang, 2024). Specifically, we test the effect of the Question Under Discussion (QUD, Roberts 2012) on inference calculation, embedding sentences such as *The movie is good* in dialogue contexts that contain polar questions like *Is the movie excellent?*. Experiment 2 also looks at how degree estimates change when the tested sentences include the focus particle *only*, e.g., *The movie is only good*. To 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. Namely, while in the inference

[2] This task shares some properties with the magnitude estimation task of linguistic acceptability (i.a., Bard et al., 1996; Fukuda et al., 2012; Loock and Auran, 2014), where participants are asked to evaluate an initial stimulus (or reference sentence), compare subsequent stimuli to it, and assign them a numerical value that reflects their relative acceptability. For instance, if a reference sentence is assigned 50, and a subsequent sentence is deemed to be twice as acceptable, it would receive a score of 100. Though both the magnitude estimation and the degree estimate task allow researchers to collect fine-grained judgments, an important difference between them is that the latter does not ask for an explicit comparison of linguistic stimuli, but instead asks participants to map sentences onto an underlying degree scale.

https://doi.org/10.1017/langcog.2024.55 Published online by Cambridge University Press

task, *only* results in more robust calculation of upper-bounded (*not excellent*) inferences than the QUD, in Experiment 2 we find that degree estimates remain comparatively higher with *only*. We provide an explanation for this discrepancy in terms of the inference task presenting participants with the stronger alternative. Finally, in the General discussion (Section 5), we compare and summarize the respective advantages and disadvantages of the inference task versus the degree estimate task.

2. Experiment 1

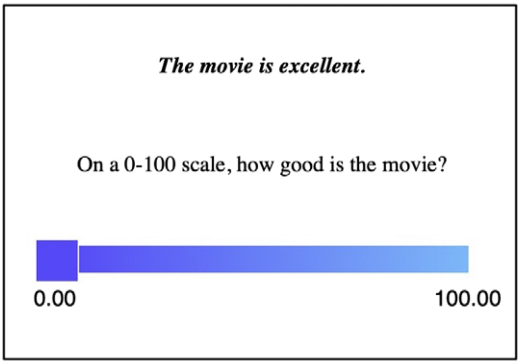
In order to validate our methodology, we first used the degree estimate task to compare weaker scalar terms to their stronger alternatives, since we had a clear prediction that the former would lead to lower degrees than the latter. Additionally, given that the possibility of negative strengthening is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

2.1. Participants, task and materials

Ninety-one 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. The experiment took an average of approximately 10 minutes to complete; thus, the compensation was in line with the Illinois minimum wage at the time. 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.

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) – that no participant saw weaker and stronger statements together will become important in the analysis reported in Section 3.1. 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 stronger alternative condition.

We aimed to create neutral task questions that would not bias participants toward either end of the scale. For adjectival lexical scales, questions relied on the weaker term wherever possible (*On a 0–100 scale, how old is the house?* for < *old, ancient* >), while in other cases we picked a neutral underlying adjective, e.g., *On a 0–100 scale, how likely is success?* for < *possible, certain* >. Questions for verbal and adverbial scales were necessarily more varied but aimed to be neutral and refer to the underlying scale, e.g., *On a 0–100 scale, how much will the sales increase?* for < *double, triple* > or *On a 0–100 scale, how often is the lawyer early?* for <*usually, always*>. It is worth addressing whether the different types of task questions lead to differences in results. For example, it is possible that mentioning the weaker scalar (e.g., *good*) in the



The movie is excellent.

On a 0-100 scale, how good is the movie?

Figure 2. Example experimental trial from Experiment 1: stronger alternative condition.

task question made the target lexical scale (< *good*, *excellent* >) salient to participants, which could have impacted their responses. To probe this possibility, in the analyses of Experiments 1–2, we will compare items where the task question relied on the weaker term ($N = 37$) with items where a different task question was used ($N = 23$). As we will see in Sections 2.3 and 4.3, the overall results – i.e., the relative order of degree estimates produced by the three conditions within each experiment – did not differ in the two sets of items. This suggests that our selection of task questions did not introduce unwanted biases into the data.

The experiment included 60 critical items, that is, 60 different lexical scales (adjectives, verbs, adverbs, quantifiers, and connectives). To identify these 60 lexical scales, we conducted the following searches in the Corpus of Contemporary American English (COCA, Davies 2008): *X or even Y*; *not just X but Y*; *X but not Y*, which have been used by van Tiel et al. (2016) and Pankratz and van Tiel (2021). We searched for adjectives, verbs, and adverbs. The expectation is that these searches would largely uncover sentences from the corpus where a lexical scale was produced – in particular, scales where *X* is the weaker scalar term and *Y* is the stronger scalar term. Sentences where *X* and *Y* were clearly not in a logical scale-mate relation (e.g., *unreasonable or even bloodthirsty*) were discarded based on researcher intuition. We took the items resulting from corpus searches and combined them with scales used in van Tiel et al. (2016) and de Marneffe and Tonhauser (2019). This resulted in a total number of 101 items. As the next step, the below semantic tests were conducted to probe whether *X* and *Y* indeed form a scale. ‘Yes’ vs. ‘No’ judgments were made by the first author and a native speaker consultant.

- Is *X and even Y* odd? – Expected answer: No
- Is *X but not Y* contradictory? – Expected answer: No
- Is *Y but not X* contradictory? – Expected answer: Yes

The *and even* test is for cancellability: if the *not Y* inference arising from *X* is an SI, it should be cancellable, that is, *Y* should be assertable (Grice, 1967). The *but not* tests probe for asymmetric entailment (Horn, 1972): *Y* should entail *X*, but not vice versa, for *X* and *Y* to qualify as scale-mates. Wherever a pair did not produce the expected ‘Yes’ or ‘No’ answer, it was excluded. Lastly, wherever one word participated in more than one scale, one of those scales was excluded, e.g., because *exclusively* occurred in both the < *primarily, exclusively* > scale and the < *mostly, exclusively* > scale, the latter was removed. This was done to prevent participants from having to respond to a particular target SI (*not exclusively*) in more than one trial.

The scalar terms appeared in carrier sentences (e.g., *The movie is good* for < *good, excellent* >) that were either adopted from previous work or generated with the goal of being natural and neutral. In addition to critical items, 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 movie 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.*[3]

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* ends up meaning that the movie is less than good. As mentioned, previous experimental work has shown that participants are indeed sensitive to negative strengthening. In Gotzner et al.'s (2018) study, for example, participants saw sentences such as *He is not brilliant* and were asked whether they could 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

Experimental data for both experiments, as well as the scripts used for data visualization and analysis can be found in the following OSF repository:

https://osf.io/fz4du/?view_only=bc7ed922a72c4cf1b7153ad67814dbac

Figure 3 shows the results of Experiment 1 as violin plots.[4] For the statistical analysis, we fit a Bayesian mixed effects zero–one-inflated beta (ZOIB) regression model using the `brms` package (Bürkner, 2017) in R (R Core Team, 2021). This type of model takes into consideration that the scale is bounded at 0 and 1 while values

[3] This interpretation is complicated by the fact that *good* and *excellent* denote intervals, yet in the degree estimate task we ask participants to pick a single point. We discuss this issue in more detail in the General discussion.

[4] Visualization of the by-scale data can be found in the Appendix: see Figure A1 for Experiment 1 and Figure A2 for Experiment 2.

https://doi.org/10.1017/langcog.2024.55 Published online by Cambridge University Press

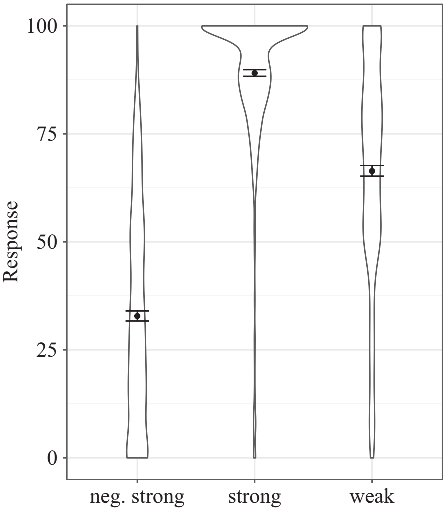


Figure 3. Experiment 1 results. Dots represent means and error bars 95% confidence intervals.

between these are modeled as a beta distribution.[5] The model predicted Response (transformed to 0–1 from the original 0–100 sliding scale ratings) by Condition (weak versus strong versus negated strong). For our analyses, we use functions from the `contrastable` R package (Sostarics, 2024). Condition was treatment-coded using the `treatment_code()` function, setting weak as the reference level. We used weakly informative priors[6] on the Condition effect of interest and the default priors for the ZOIB distributional parameters. The random effects structure included by-participant random intercepts and by-item random intercepts and slopes. We report the posterior parameter estimates $(\hat{\beta})$ with the 95% credible intervals (CI). We consider the CI excluding zero to be evidence for an effect. The analysis revealed that Responses to strong terms were higher than to weak terms $(\hat{\beta} = 1.04, \mathrm{CI} : [0.75, 1.34])$. Responses to negated strong terms were lower than to weak terms $(\hat{\beta} = -1.07, \mathrm{CI} : [-1.36, -0.78])$. Figure 4 shows the posterior predicted means from the model for each condition.

Next, to probe whether items (N = 37) where the task question contained the weaker scalar term differed from items (N = 23) where it did not, we conducted an additional analysis (see Section 2.1 for the motivation). For this, we set up an additional Question predictor, with the level ‘same’ for the former set of items and ‘different’ for the latter set. We then fit a Bayesian mixed effects ZOIB regression model predicting Response (0–1) by Condition (weak versus strong versus negated strong), Question (same versus different) and their interaction, with weakly informative priors on Condition and Question. The random effects structure included by-participant and by-item random intercepts, as well as random slopes of Question by participant and random slopes of Condition by items. Condition is again

[5] We thank an anonymous reviewer for suggesting this model.

[6] We also tried additional priors that were more in line with the direction of the Condition effects, as well as the default priors for all parameters, but neither of these affected the results.

https://doi.org/10.1017/langcog.2024.55 Published online by Cambridge University Press

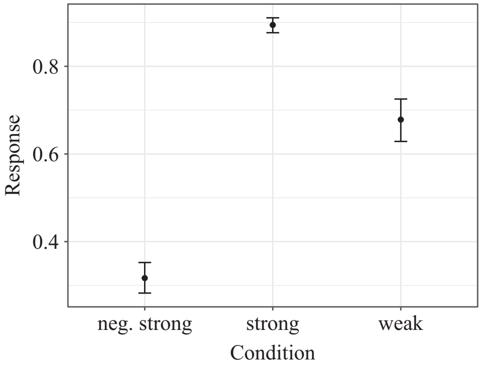


Figure 4. Experiment 1 posterior predicted means from the Bayesian mixed effects ZOIB regression model. Error bars show 95% credible intervals.

treatment coded, setting weak as the reference level, and Question is coded using the `scaled_sum_code()` function (+.5/−.5), setting same as the reference level. Using these contrasts, the fixed effects of Condition now average over the two question sets. We find that the patterns reported above still hold for both the strong versus weak comparison $\left(\hat{\beta}=1.06, \mathrm{CI}:[0.77,1.34]\right)$ and the negated strong versus weak comparison $\left(\hat{\beta}=-1, \mathrm{CI}:[-1.27,-0.73]\right)$. Though there are credible interactions between Question and Condition (negated strong: $\hat{\beta}=0.98, \mathrm{CI}:[0.57,1.39]$; strong: $\hat{\beta}=0.57, \mathrm{CI}:[0.12,1.01]$), these primarily reflect how Question only made a difference for our reference level, i.e., the weak condition. This pattern is shown in Figure 5. As can also be seen in that figure, the negated strong < weak < strong pattern holds for both Question types. In sum, variations in the task question did not have considerable impact on the overall patterns.

Our first finding was that, when averaging 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

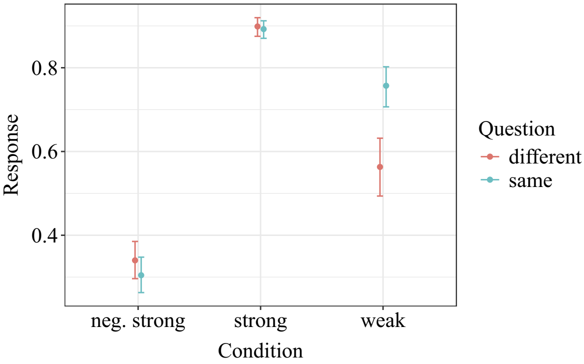


Figure 5. Experiment 1 posterior predicted means from the Bayesian mixed effects ZOIB regression model with the additional Question predictor. Error bars show 95% credible intervals.

https://doi.org/10.1017/langcog.2024.55 Published online by Cambridge University Press

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.* Otherwise, given that the non-SI-enriched meaning of *good* is compatible with *excellent*, we might not have expected a difference between these two conditions. 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 not only that the movie is less than excellent (predicted by the semantic contribution of negation), but that the movie is in fact less than good. This can be interpreted as negative strengthening (Horn, 1989), confirming that our experimental paradigm is able to detect such pragmatic inferences.

3. Correlations with likelihood of SI calculation

In Experiment 1, we used the degree estimate task to test sentences containing the weaker scalar term, its stronger alternative, as well as the negated version of the stronger alternative. Our results averaged over different lexical scales served as a reality check: all three conditions were different from each other, with the strong terms being highest and the negated strong terms being lowest (due to negative strengthening). In this section, we continue to explore the Experiment 1 data, this time looking at by-scale variation. First, we employ degree estimates to further test van Tiel et al.'s (2016) distinctness hypothesis (Section 3.1). Successfully replicating that distinctness correlates with the likelihood of SI calculation provides another reality check on the degree estimate task. Second, we find that the degree estimates of the negated stronger term are also a significant predictor of across-scale variability in inference task results (Section 3.2). However, since the inference task directly mentions the negated stronger alternative, this finding is likely an artifact of participants being presented with that expression.

3.1. Distinctness as a predictor of scalar diversity

As mentioned in Section 1.1, distinctness of scalar terms was originally put forth by van Tiel et al. (2016) as a potential explanation for scalar diversity. Distinctness is relevant for the likelihood of SI calculation for the following reason. The inferential process underlying SI calculation involves the hearer reasoning about and negating a stronger alternative (*all*) that the speaker could have said but did not. For this reasoning to go through, there has to be a clear stronger alternative, and it has to be sufficiently stronger. In other words, the more distinct two scalar terms (*some* versus *all*) are, the more likely the hearer is to assume that the speaker should have used the stronger term if possible. If it is difficult to distinguish the weak and strong scalar, e.g., if they are near-synonyms, SI calculation is unlikely. Here, we use degree estimates to operationalize distinctness by comparing the degree estimate given to a weaker term from a scale (e.g., *good*) to that given to the stronger alternative to that term (*excellent*). We can predict that the greater the difference between the degree estimates for the weak and the strong scalar terms, i.e., the further apart they are on the underlying degree scale, the higher the SI rate will be for that scale. As mentioned, this is because for an SI (*good but not excellent*) to

https://doi.org/10.1017/langcog.2024.55 Published online by Cambridge University Press

arise, *good* and *excellent* have to be perceived as describing two different world states.

To check the prediction of distinctness, relying on the Experiment 1 data, we took the absolute difference in means between the weak and strong terms for every lexical scale. For instance, *The movie is good* received a response of 69.4 on the 0–100 scale, while *The movie is excellent* received 89.1, resulting in a 'distinctness' value of 19.7. In order to see whether distinctness significantly predicts the likelihood of SI calculation, we correlated the obtained values with the rate of SI calculation. The SI rates were taken from Ronai and Xiang (2024, Experiment 1), who used the inference task to measure SI calculation from the same 60 lexical scales we test here. Figure 6 shows these results. As can be seen in the figure, there was a positive correlation between the degree estimate-based distinctness values and SI rates (Pearson's correlation test: $r = 0.33, p < 0.05$). That is, scalar diversity was shown to be predicted by the distinctness of scalemates. Specifically, the higher the degree estimate difference between a weak (*good*) and a strong (*excellent*) term, the higher the corresponding SI rate from that scale (*good but not excellent*).

In other words, we found that the more distinct the world states that the weaker and the stronger term on a scale are taken to describe, the higher the SI rate for that scale. These results thus present evidence for van Tiel et al.'s (2016) distinctness hypothesis, using a novel operationalization that relies on empirically collected degree estimates. Van Tiel et al. relied on the notion of boundedness, as well as experimentally collected judgements about semantic distance, to test the distinctness hypothesis. It is worth discussing how the latter relates to our findings. In van Tiel et al.'s semantic distance experiment, participants were presented with a pair of sentences, such as *She is intelligent* and *She is brilliant*. They then had to respond to

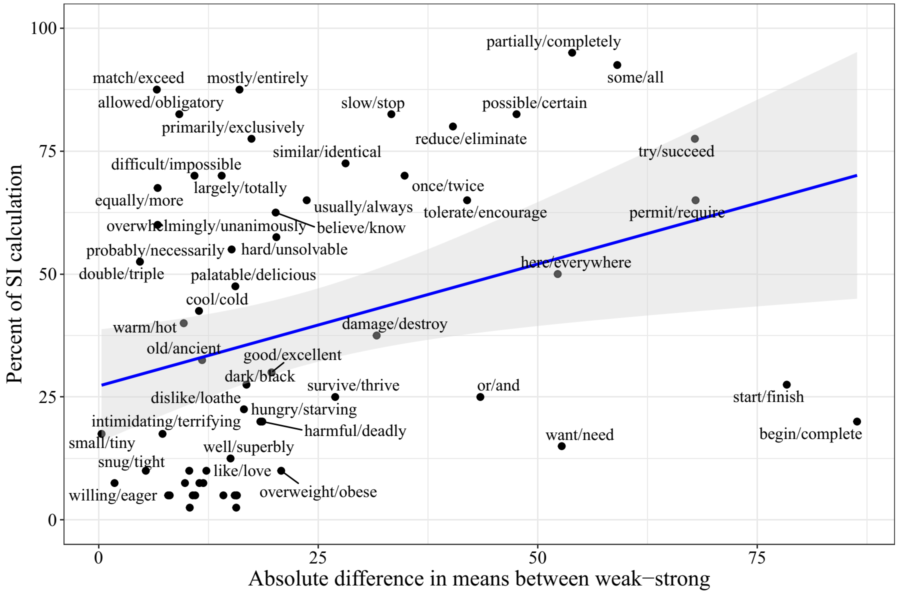


Figure 6. The *x*-axis shows distinctness between each weak-strong scalar pair from Experiment 1. The *y*-axis shows SI rates from Ronai and Xiang (2024, Experiment 1).

https://doi.org/10.1017/langcog.2024.55 Published online by Cambridge University Press

the question *Is statement 2 stronger than statement 1?* on a 7-point Likert scale, where 1 corresponded to ‘equally strong’ and 7 to ‘much stronger’. In line with the distinctness hypothesis, the authors found that semantic distance was positively correlated with SI rates: the more distant a weak and a strong scalar term were in their experiment, that is, the stronger statement 2 was judged to be, the more likely the corresponding SI.

Operationalizing distinctness using degree estimates has advantages for the following reasons. First, van Tiel et al.’s semantic distance experiment assumed an *a priori* strength relation: statement 2 had to be minimally as strong as statement 1. On the contrary, our experimental instructions did not presuppose that one statement could be stronger than the other (or which one that would be), and participants simply identified degrees (corresponding to the weaker versus stronger statement) on an underlying scale. Second, van Tiel et al. presented the weaker and stronger statements together on the same screen. Being explicitly provided with the stronger alternative statement (*She is brilliant*) could have encouraged participants to calculate the SI from the weaker statement (*She is intelligent*). In case SI calculation did happen, then what participants were ultimately judging is not whether (and by how much) *She is brilliant* is a stronger statement than *She is intelligent*, but instead how much stronger *She is brilliant* is than *She is intelligent but not brilliant.* If this is the case, then the results obtained do not necessarily reflect the semantic distance between (unstrengthened) *intelligent* and *brilliant.* Recall that in our experiments, the strong versus weak manipulation was conducted between participants, so no participant saw both the weaker and stronger statement. Therefore, while it is still possible that SIs were calculated from the weaker statements and influenced the obtained distinctness values, there was no built-in bias to encourage this. Third and finally, judging the relative strength of statements (as in van Tiel et al. 2016) requires a metalinguistic judgment, while providing degree estimates is arguably a more natural task. Altogether, replicating the distinctness finding using degree estimates constitutes further evidence for van Tiel et al.’s hypothesis, going beyond existing evidence in the prior literature, and it also further validates the degree estimate task itself.

3.2. The negated strong scalar as a predictor

Many of the previously identified predictors of scalar diversity concern the relationship between the weak and the strong scalar term, e.g., their semantic distance, semantic similarity, or the availability of the stronger given the weaker. But since the inference task contains the negated version of the stronger alternative, which allows for non-SI inferences like negative strengthening to impact judgments, it is also conceivable that the meaning of the *negated* alternative (e.g., *not excellent*) could be a predictor. As we saw, the Experiment 1 findings, averaged over all lexical scales, suggest that the degree estimate task is successful at detecting negative strengthening. In the following analysis, we look at the by-scale variation in this data, probing whether the meaning of the negated strong term (*not excellent*) plays a role in the variation in inference task results.

Suppose, for instance, that *good* and *not excellent* are interpreted as describing two very different world states – that is, they are distant on the degree scale of goodness. Now consider the inference task, which asks whether the participant can conclude

https://doi.org/10.1017/langcog.2024.55 Published online by Cambridge University Press

from Mary's utterance *The movie is good* that she thinks the movie is not excellent. If *good* and *not excellent* indeed describe very different world states, then 'No' responses are more likely – the participant reasons that Mary treats *good* and *not excellent* as different – and this result is interpreted as a low SI rate. On the other hand, if *good* and *not excellent* represent similar world states, then 'Yes' responses might be more likely if the participant reasons that Mary considers *good* and *not excellent* to be near equivalent, and this result looks like a high SI rate. Our prediction is that the smaller the difference between the degree estimates for the weak and the negated strong term, the higher the rate of 'Yes' responses will be. In other words, we predict a negative correlation between the weak-negated strong degree estimate difference and rate of 'Yes' responses.

To check the prediction that the meaning of the negated strong term captures across-scale variation in the rate of inference task 'Yes' responses, we again calculated the absolute difference in means for every scale in Experiment 1: this time between the response to the weak term and the response to the negated strong term. For example, for the < *good, excellent* > scale, *The movie is good* received a response of 69.4 on the 0–100 scale, while *The movie is not excellent* received 31.5, resulting in a score of 37.9 – these are plotted on the *x*-axis of Figure 7. There was a negative correlation between these results and the SI rates from Ronai and Xiang (2024, Experiment 1) (Pearson's correlation test: $r = -0.61, p < 0.001$). In other words, we found that the more similar the world states that a weaker and negated stronger term are taken to describe, the higher the rate of 'Yes' responses for that scale.

The motivation for the current analysis was that the inference task commonly used to test SI calculation explicitly mentions the negated stronger term, which raises the

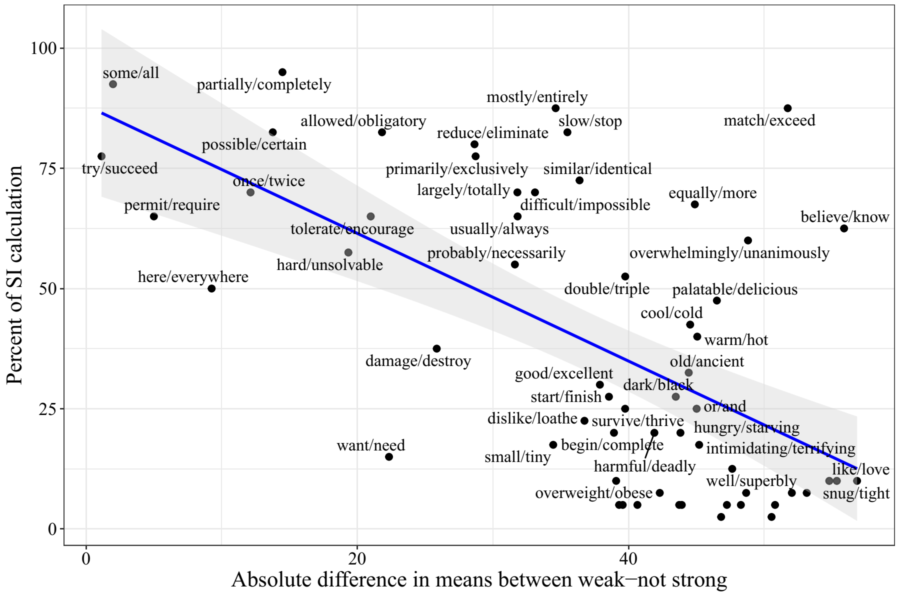


Figure 7. The *x*-axis shows the meaning of the negated stronger term from Experiment 1. The *y*-axis shows SI rates from Ronai and Xiang (2024, Experiment 1).

possibility that, when participants choose not to endorse the conclusion that a speaker meant *not excellent* by uttering *good*, they do so because they perceive *not excellent* and *good* to mean different things. Our findings suggest that the meaning of the negated strong term (vis-à-vis the meaning of the weak term), as measured by experimentally collected degree estimates, indeed captures some of the variation in inference task results that is observed across scales. That the rate of 'Yes' responses in the inference task is related to the similarity in meaning between the weak and the negated stronger term supports the possibility that participants may sometimes respond with 'No' not due to a lack of SI calculation but as an artifact of being presented with the negated strong term, which they perceive to mean something different from the weak term. As such, the inference task's task question explicitly mentioning the negated stronger alternative can be viewed as a limitation since it might create an illusion of SI non-calculation.

4. Experiment 2

In Experiment 2, we use the degree estimate task to reassess previous findings from experimental work that used the inference task. Specifically, we turn to two manipulations that are known to increase the likelihood of upper-bounded inference calculation. Both manipulations of interest relate to the assumption that inferences like *good but not excellent* arise via hearers' reasoning about unsaid alternatives. In what follows, alternatives are made salient in two different ways. First, prior to the target sentence *The movie is good*, we add a question that explicitly probes whether the movie is excellent. This introduces a discourse context that encourages participants to consider the stronger alternative utterance *The movie is excellent.* Second, in a different condition, we add the focus particle *only* to the target sentences (e.g., *The movie is only good*). The focus particle semantically encodes alternative exclusion in the asserted content, again engaging participants in actively considering alternatives. Previous work by Ronai and Xiang (2024) has employed these two manipulations in the inference task. By conducting these manipulations in the degree estimate paradigm, we are able to test whether previous inference task-based results still hold.

4.1. Participants, task and materials

Ninety-seven native speakers of American English participated in an experiment on the Ibex platform for either $2 (*only* experiment) or $2.25 (QUD experiment) compensation. The experiments took on average 10–12 minutes to complete; thus the compensation was in line with the Illinois minimum wage at the time. Participant recruitment and screening were 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) or the weaker scalar term itself (4).

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

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

The inference-triggering sentences were modified to ensure dialogue coherence, e.g., Mary's utterance of *The movie is good* was changed to *It is good.* Otherwise, the sentences were identical to Experiment 1. The weak versus strong QUD manipulation was administered within participants. The key expectation for the effect of this manipulation is that in the strong QUD condition, the likelihood of SI calculation should increase. One reason this would arise is that Sue's question explicitly mentions the stronger alternative. When Mary gives an answer that declines to use the stronger alternative made salient by the question, this encourages participants' reasoning about that alternative and consequently calculating the SI. This expected effect can also be captured by theoretical proposals such as the Question-Answer Congruence account of Hulsey et al. (2004), and is in line with previous experimental findings from i.a., Yang et al. (2018) and Zondervan et al. (2008).

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

(5) The movie is only good. (*only* condition)

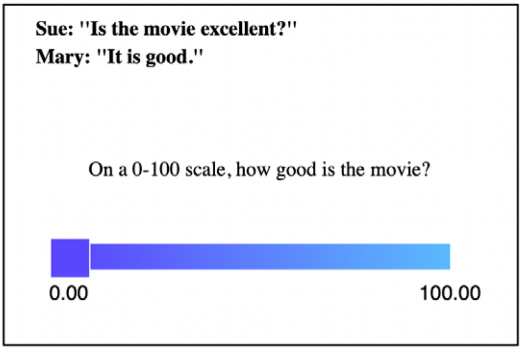
As mentioned, the focus particle *only* encodes the exclusion of alternatives semantically (Rooth, 1985, 1992). Similarly to the strong QUD manipulation, this encourages participants to consider alternatives and is expected to increase the likelihood of calculating the *good but not excellent* inference. Additionally, a stronger effect is expected from the *only* manipulation than from the QUD manipulation, since in the case of *only*, the upper-bounded meaning is no longer a cancellable pragmatic inference, but is instead encoded in the asserted content.

Experiment 2 used the same 60 lexical scales as critical items as Experiment 1 did. The *only* experiment had 60 critical items in a single condition, while the QUD experiment contained the same 60 items in 2 within-participant conditions. In addition to 60 critical items, each experiment also included 3 practice and 5 filler items. Practice and filler items were slightly modified from Experiment 1: e.g., to better serve as catch trials (*The table is clean* → *The table is 100% clean*), and in the QUD experiment, they now included an explicit question to match the critical items (e.g., Sue: *Is the table clean?*; Mary: *It is 100% clean.*). Experiment 2 was otherwise identical to Experiment 1 in its instructions, task questions ('On a 0–100 scale, how good is the movie?') and procedure. Figure 8 shows an example trial from the strong QUD condition.

4.2. Hypotheses and predictions

For our predictions, let us turn to findings from Ronai and Xiang (2024), who used the inference task to test conditions identical to the current Experiment 2 (sentences like (3)–(5)). First, Ronai and Xiang (2024) reported 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 (4). Second, the

https://doi.org/10.1017/langcog.2024.55 Published online by Cambridge University Press



Sue: "Is the movie excellent?"
Mary: "It is good."

On a 0-100 scale, how good is the movie?

Figure 8. Example experimental trial from Experiment 2: strong QUD condition.

presence of the focus particle *only* (5) resulted in inference rates even higher than the supportive context of the strong QUD condition in (3).

Overall, if the degree estimate task were to replicate findings from the inference task, we would expect to find the most robust inference calculation in the *only* condition, followed by the strong QUD condition, with the weak QUD condition leading to the least inference calculation. In the degree estimate task, a higher rate of upper-bounded inference calculation corresponds to a lower degree estimate of the target property: the more robust the calculation of a *good but not excellent* inference from *good*, the lower the degree of goodness attributed to the statement *The movie is (only) good.* For the three conditions in (3) to (5), therefore, we would expect that the *only* condition will result in the lowest degree estimates, the strong QUD condition in the second lowest degree estimates, and the weak QUD condition will have the highest degree estimates.

4.3. Results and discussion

Figure 9 shows the results of Experiment 2 as violin plots. For the statistical analysis, we fit a Bayesian mixed effects ZOIB regression model predicting Response (0–1) by Condition (weak QUD versus strong QUD versus *only*), with weakly informative priors on Condition. Condition was again treatment coded, setting weak QUD as the reference level. The random effects structure included by-participant random intercepts and by-item random intercepts and slopes. As compared to the weak QUD condition, the analysis found lower Responses for both the *only* $\left(\hat{\beta} = -0.29, \mathrm{CI}: [-0.52, -0.05]\right)$ and strong QUD $(\hat{\beta} = -0.47, \mathrm{CI}: [-0.55, -0.38])$ conditions. For an additional pair comparison, we also fit a model where the *only* condition served as the reference level (the model was otherwise identical). While the effect for strong QUD (as compared to *only*) contains 0 in the 95% credible interval $(\hat{\beta} = -0.17, \mathrm{CI}: [-0.41, 0.06])$, the probability of direction (here, negative) is 92.4%. Additionally, this effect only reflects the values between (not including) 0 and 1. When considering the number of extreme values (a 0 or 1), we find that the strong QUD condition includes fewer extreme values $(\hat{\beta} = -1.18, \mathrm{CI}: [-1.4, -0.97])$ and, in particular, much fewer 1s $(\hat{\beta} = -1.34, \mathrm{CI}: [-1.91, -0.76])$ than *only*. (The strong QUD condition having fewer 100 responses than *only* can be seen in Figure 9.) In total, the model predicted proportion

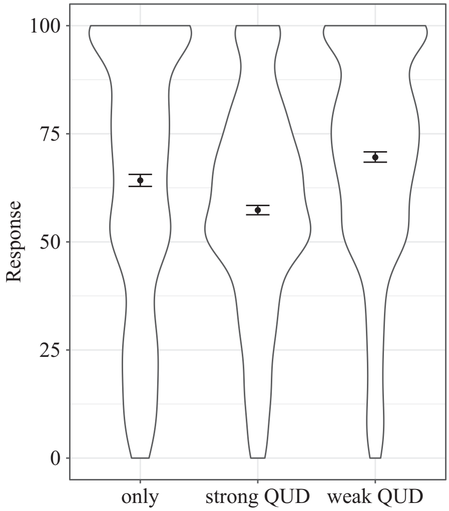


Figure 9. Experiment 2 results. Dots represent means and error bars 95% confidence intervals.

of 1s is 17% in the *only* condition but only 5% in the strong QUD condition. Taken together, these results suggest that the strong QUD condition has lower Responses overall than the *only* condition. Figure 10 shows the posterior predicted means from the model for each condition.

Similarly to Experiment 1, to rule out the possibility that the task question sometimes containing the weaker scalar term could have impacted the results, we conducted an additional analysis. We fit a Bayesian mixed effects ZOIB regression model predicting Response (0–1) by Condition (weak QUD versus strong QUD versus *only*), Question (same: when it contained the weaker scalar versus different: when it did not) and their interaction, with weakly informative priors on Condition and Question. Condition was again treatment coded, with weak QUD as the reference level, and Question was again scaled sum coded, with same as the reference level. The random effects structure included by-participant and by-item random intercepts, as well as random slopes of Question by

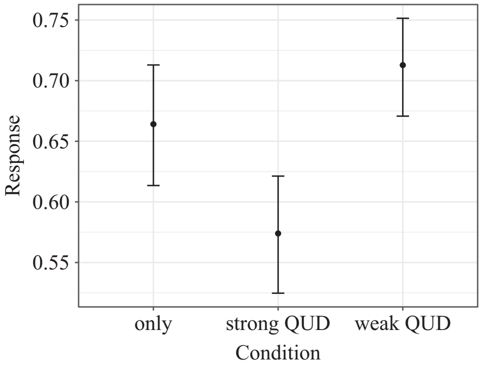


Figure 10. Experiment 2 posterior predicted means from the Bayesian mixed effects ZOIB regression model. Error bars show 95% credible intervals.

https://doi.org/10.1017/langcog.2024.55 Published online by Cambridge University Press

participant and random slopes of Condition by items. The model finds no credible interaction between Condition and Question (*only*: $\hat{\beta} = 0.02, \mathrm{CI}: [-0.24, 0.27]$; strong QUD: $\hat{\beta} = 0.01, \mathrm{CI}: [-0.17, 0.18]$). Crucially, the fixed effects of Condition still show the pattern reported above, with *only* $(\hat{\beta} = -0.28, \mathrm{CI}: [-0.50, -0.06])$ and strong QUD $(\hat{\beta} = -0.48, \mathrm{CI}: [-0.56, -0.39])$ both producing lower Responses than weak QUD. We then repeated this analysis but set *only* as the reference level. Again, we find no credible interaction between Condition and Question (strong QUD: $\hat{\beta} = -0.01, \mathrm{CI}: [-0.28, 0.25]$; weak QUD: $\hat{\beta} = -0.02, \mathrm{CI}: [-0.28, 0.24]$). The effect of strong QUD being lower than *only* $(\hat{\beta} = -0.19, \mathrm{CI}: [-0.41, 0.04])$ also patterns similarly to the analysis in the previous paragraph, with a probability of direction of 94.9%. Thus, we can conclude that the task question formulation did not impact the results of Experiment 2.

To summarize, we found that the degree of goodness attributed to the movie was highest in the weak QUD condition (see (4)). The degree estimate was lower for the *only* condition in (5) and the lowest for the strong QUD condition in (3). This is not a replication of the findings obtained by Ronai and Xiang (2024). Concretely, using the inference task, Ronai and Xiang (2024) found more *good but not excellent* inferences with *only* than with the strong QUD. Yet using the degree estimate task, the present Experiment 2 found lower degrees of goodness with the strong QUD than with *only* – where we take lower degrees to correspond to a more robust calculation of *good but not excellent* inferences. That is to say, the current findings suggest more upper-bounded inference calculation in the strong QUD condition than the *only* condition.

A potential explanation for the differences between the current experiment and the findings in Ronai and Xiang (2024) is as follows. 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 exclusion of non-scalar alternatives, e.g., that the movie is not funny (i.a., Coppock and Beaver, 2013). 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*). The possibility of excluding non-scalar alternatives under *only* might have allowed degree estimates (on e.g., the degree scale of goodness) in the *only* condition to remain higher. In the inference task, on the other hand, we have noted that the task question ('Would you conclude from this that Mary thinks the movie is not excellent?') explicitly mentions the stronger scalar alternative *excellent*, which biases participants to reason about that particular alternative. Thus, when participants are provided with alternatives like *excellent*, they are more likely to take those to be the ones excluded by *only* than when they are merely presented with the inference-triggering sentences – the latter situation better reflects how upper-bounded inferences arise in real-life communication. Being provided with specific alternatives then has the result of inflating the rates of calculating the *good but not excellent* meaning and the corresponding 'Yes' responses in the inference task.

In sum, in Experiment 2 we found lower degree estimates for the strong QUD condition than the *only* condition, which is counter to what has been found in previous work using the inference task. We have proposed an explanation for this discrepancy with reference to the different properties of the two tasks.

https://doi.org/10.1017/langcog.2024.55 Published online by Cambridge University Press

5. General discussion

An inference task is often used to test SI calculation; it is especially common in investigations of scalar diversity. However, such a task is biasing, as it provides participants with a particular stronger scalar alternative; it also obscures whether other, non-SI inferences factor in participants' 'Yes' vs. 'No' responses. In this article, we introduced an alternative experimental task that relied on collecting degree estimates: asking participants to judge on an underlying scale, e.g., how good a movie is, given that it has been asserted that *The movie is good.* We validated the degree estimate task by first comparing judgements to weaker scalar terms (*good*), their stronger alternatives (*excellent*), and the negated version of those alternatives (*not excellent*) (Experiment 1). The findings of this experiment largely served as a reality check and confirmed that participants performed the task adequately. But we were also able to operationalize the distinctness of scalemates using degree estimates and showed that this correlates with inter-scale variation in SI rates. We argued that this replicates, i.a., van Tiel et al.'s (2016) findings, using a task that has advantages over their Likert scale-based measure of semantic distance. Crucially, another analysis of the by-item results showed that the meaning of the negated stronger term can predict rates of 'Yes' responses in the inference task – this can be interpreted as evidence that explicitly mentioning this negated alternative constitutes a limitation of that task.

Finally, we also used the degree estimate task to test the role of a supportive context and *only* in modulating upper-bounded inference calculation (Experiment 2). Our results were not entirely in line with previous work that used the inference task (Ronai and Xiang, 2024). Concretely, Ronai and Xiang's inference task findings revealed that supportive contexts lead to an increase in SI calculation rates. But in that study, the likelihood of calculating inferences such as *good but not excellent* was in fact highest when sentences included the focus particle *only*. In the current study 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 weak QUD condition. We explained the discrepancy between the two sets of results by reference to a relevant feature of the inference task that was one of the starting points of our paper, namely the bias created by the mention of the stronger alternative. Altogether, the results of our experiments highlight the value of using a more fine-grained, rather than binary, experimental measure in the study of SI and scalar diversity.

Although we have shown that the degree estimate task can avoid the disadvantages of the traditional inference task and yield theoretically informative results, there are some limitations of the task that call for more future work. We have used the degree estimate task to probe the interpretation of utterances containing scalar terms by assuming that lower estimates correspond to more upper-bounded meaning. But 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. Here it is important to note that from a semantic perspective, gradable predicates like *good* and *excellent* relate individuals to degree intervals on a scale (Kennedy and McNally, 2005, among many others). If a participant computes the SI interpretation *good but not excellent*, they would represent the two predicates as referring to two non-overlapping degree intervals, with the interval for *excellent* on the higher end of the 'goodness' scale than the interval for *good*. But if a participant does not compute the SI, the two intervals would overlap, with the onset of the *good* interval starting from a lower degree threshold than the onset of the *excellent* interval. In the degree estimate task, however, we only asked participants to make a point estimate, instead of an interval estimate. It is possible that comparing the averaged point estimates between *good* and *excellent* is not an ideal proxy for comparing two degree intervals. For example, if a participant adopts a response strategy to always choose the middle point of an underlying degree interval, this would yield a lower degree estimate for *good* than *excellent* even when the two intervals are overlapping, creating an illusion of SI calculation. As a result, when we observe a difference between the degree estimates of a pair of scalemates, we could not be absolutely certain that an SI has been generated.

One could argue that the above caveat may be addressed by collecting 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 importantly, the *not excellent* meaning is now part of the asserted content. Even though *The movie is good but not excellent* describes the same world state that hearers would arrive at having calculated the SI from *The movie is good*, it is possible that the degree estimate results would differ when *not excellent* is in the semantics and therefore has a very different status from an SI. Therefore, this option also does not represent an unproblematic candidate for tapping into SI calculation directly via the degree estimate task. Another alternative would be to complement the degree estimate task with the traditional inference task, since the latter task explicitly probes for the SIs. But future work should also explore alternative experimental tasks or data analysis techniques that could better capture the interval degree semantics.

Summing up, in this article, we have highlighted some of the shortcomings of the inference task: namely, that it makes the stronger alternative explicit and that it obscures additional pragmatic inferences (see also i.a., Geurts and Pouscoulous 2009 and Benz et al. 2018 for these observations). We explored the possibility of studying scalar meanings with the novel degree estimate task instead. But as noted, the degree estimate task is not without shortcomings either, as it collects a point estimate and does not provide a straightforward metric of when SI calculation has occurred. Altogether, each task clearly has its own set of advantages and disadvantages; reassuringly, we have also seen that they can both identify some core findings about scalar diversity, namely semantic distance effects. Given the inference task's virtual monopoly in the study of scalar diversity, we argue that degree estimates represent an interesting new way of looking at SI and scalar diversity.

Data availability statement. Stimuli, data, and the scripts used for data visualization and analysis can be found in the following OSF repository: https://osf.io/fz4du/?view_only=bc7ed922a72c4cf1b7153ad67814dbac

Acknowledgments. We would like to thank Larry Horn, Chris Kennedy, Hannah Rohde, Michael Tabatowski, and two anonymous reviewers for helpful discussion and feedback on this project. We are also grateful to Sherry Yong Chen for help with the implementation of the sliding scale task and Thomas Sostarics for help with the ZOIB statistical models.

Funding statement. This material is partially based upon work supported by the National Science Foundation under Grant No. #BCS-2041312.

References

Baker, R., Doran, R., McNabb, Y., Larson, M., & Ward, G. (2009). On the non-unified nature of scalar implicature: An empirical investigation. *International Review of Pragmatics*, 1(2), 211–248.

Bard, E. G., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, 72(1), 32–68.

Beltrama, A., & Xiang, M. (2013). Is 'good' better than 'excellent'? An experimental investigation on scalar implicatures and gradable adjectives. In E. Chemla, V. Homer, & G. Winterstein (Eds.), *Proceedings of Sinn und Bedeutung 17* (pp. 81–98). https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/333.

Benz, A., Bombi, C., & Gotzner, N. (2018). Scalar diversity and negative strengthening. In U. Sauerland and S. Solt (Eds.), *Proceedings of Sinn und Bedeutung 22* (pp. 191–203).

Breheny, R. (2008). A new look at the semantics and pragmatics of numerically quantified noun phrases. *Journal of Semantics*, 25(2), 93–139.

Bürkner, P.-C. (2017). brms: An r package for Bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1), 1–28. R package version 2.16.3.

Coppock, E., & Beaver, D. I. (2013). Principles of the exclusive muddle. *Journal of Semantics*, 31(3), 371–432.

Cummins, C., & Rohde, H. (2015). Evoking context with contrastive stress: Effects on pragmatic enrichment. *Frontiers in Psychology*, 6, 1779.

Davies, M. (2008). The Corpus of Contemporary American English (COCA). https://www.english-corpora.org/coca/ (accessd September 2020).

de Marneffe, M.-C., & Tonhauser, J. (2019). Inferring meaning from indirect answers to polar questions: The contribution of the rise-fall-rise contour. In M. Zimmermann, K. von Heusinger, & E. Onea (Eds.), *Current research in the semantics/pragmatics interface* (vol. 36, Questions in Discourse, pp. 132–163). Brill.

Degen, J. (2015). Investigating the distribution of *some* (but not *all*) implicatures using corpora and web-based methods. *Semantics and Pragmatics*, 8(11), 1–55.

Doran, R., Ward, G., Larson, M., McNabb, Y., & Baker, R. E. (2012). A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language*, 88(1), 124–154.

Drummond, A. (2007). Ibex Farm. http://spellout.net/ibexfarm (accessd September 2021).

Fukuda, S., Goodall, G., Michel, D., & Beecher, H. (2012). *Proceedings of the 29th West Coast Conference on Formal Linguistics*. Somerville, MA: Cascadilla Proceedings Project, pp. 328–336.

Geurts, B., & Pouscoulous, N. (2009). Embedded implicatures?!? *Semantics and Pragmatics*, 2(4), 1–34.

Gotzner, N., Solt, S., & Benz, A. (2018). Scalar diversity, negative strengthening, and adjectival semantics. *Frontiers in Psychology*, 9, 1659.

Grice, H. P. (1967). Logic and conversation. In P. Grice (Ed.), *Studies in the way of words* (pp. 41–58). Harvard University Press.

Hirschberg, J. B. (1985). *A theory of scalar implicature* [PhD thesis]. University of Pennsylvania.

Horn, L. R. (1972). *On the semantic properties of logical operators in English* [PhD thesis]. UCLA.

Horn, L. R. (1989). *A natural history of negation*. University of Chicago Press.

Hu, J., Levy, R., & Schuster, S. (2022). *Predicting scalar diversity with context-driven uncertainty over alternatives.* Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pp. 68–74. https://openreview.net/forum?id=rTzgPQtx8-9.

Hulsey, S., Hacquard, V., Fox, D., & Gualmini, A. (2004). The question-answer requirement and scope assignment. In A. Csirmaz, A. Gualmini, & A. Nevins (Eds.), *MIT working papers in linguistics* (pp. 71–90). MITWPL.

Jasbi, M., Waldon, B., & Degen, J. (2019). Linking hypothesis and number of response options modulate inferred scalar implicature rate. *Frontiers in Psychology*, 10. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.00189/full.

Katsos, N., & Bishop, D. V. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition*, 120(1), 67–81.

Kennedy, C., & McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language*, 81(2), 345–381.

Koenig, J.-P. (1991). *Scalar predicates and negation: Punctual semantics and interval interpretations.* Proceedings of the 27th meeting of the Chicago Linguistic Society, number Part 2, pp. 140–155. Chicago Linguistics Society, Chicago.

https://doi.org/10.1017/langcog.2024.55 Published online by Cambridge University Press

Loock, R., & Auran, C. (2014). Magnitude estimation: can it do something for your pragmatics? *Corela. Cognition, Représentation, Langage*, 12(1), 1–27.

Pankratz, E., & van Tiel, B. (2021). The role of relevance for scalar diversity: a usage-based approach. *Language and Cognition*, 13(4), 562–594.

R Core Team. (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. R package version 4.1.2.

Roberts, C. (1996/2012). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5(6), 1–69.

Ronai, E., & Xiang, M. (2021). Exploring the connection between Question Under Discussion and scalar diversity. *Proceedings of the Linguistic Society of America*, 6(1), 649–662.

Ronai, E., & Xiang, M. (2024). What could have been said? Alternatives and variability in pragmatic inferences. *Journal of Memory and Language*, 136, 104507.

Rooth, M. (1985). *Association with focus* [PhD thesis]. University of Massachusetts, Amherst.

Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics*, 1(1), 75–116.

Ruytenbeek, N., Verheyen, S., & Spector, B. (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, L., Kim, M., & Grodner, D. J. (2019). Social context modulates tolerance for pragmatic violations in binary but not graded judgments. *Frontiers in Psychology*, 10. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.00510/full.

Simons, M., & Warren, T. (2018). A closer look at strengthened readings of scalars. *Quarterly Journal of Experimental Psychology*, 71(1), 272–279.

Solt, S., & Waldon, B. (2019). Numerals under negation: Empirical findings. *Glossa: a journal of general linguistics*, 4(1), 113. https://doi.org/10.5334/gjgl.736.

Sostarics, T. (2024). *contrastable: Contrast coding utilities in R.* R package version 0.3.0. https://cran.r-project.org/package=contrastable (accessed August 2024).

Sun, C., & Breheny, R. (2022). The role of alternatives in the interpretation of scalars and numbers: Insights insights from the inference task. *Semantics and Pragmatics*, 15(8), 1–15.

Sun, C., Tian, Y., & Breheny, R. (2018). A link between local enrichment and scalar diversity. *Frontiers in Psychology*, 9. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2018.02092/full.

Sun, C., Tian, Y., & Breheny, R. (2023). A corpus-based examination of scalar diversity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 50(5), 808–818. https://doi.org/10.1037/xlm0001278.

van Tiel, B., Van Miltenburg, E., Zevakhina, N., & Geurts, B. (2016). Scalar diversity. *Journal of Semantics*, 33(1), 137–175.

Westera, M., & Boleda, G. (2020). A closer look at scalar diversity using contextualized semantic similarity. *Proceedings of Sinn und Bedeutung*, 24(2), 439–454.

Yang, X., Minai, U., & Fiorentino, R. (2018). Context-sensitivity and individual differences in the derivation of scalar implicature. *Frontiers in Psychology*, 9, 1720.

Zondervan, A., Meroni, L., & Gualmini, A. (2008). Experiments on the role of the question under discussion for ambiguity resolution and implicature computation in adults. In T. Friedman, & S. Ito (Eds.), *Proceedings of semantics and linguistic theory (SALT) 18*. Amherst: The University of Massachusetts, (pp. 765–777). https://journals.linguisticsociety.org/proceedings/index.php/SALT/issue/view/91.

https://doi.org/10.1017/langcog.2024.55 Published online by Cambridge University Press

Appendix

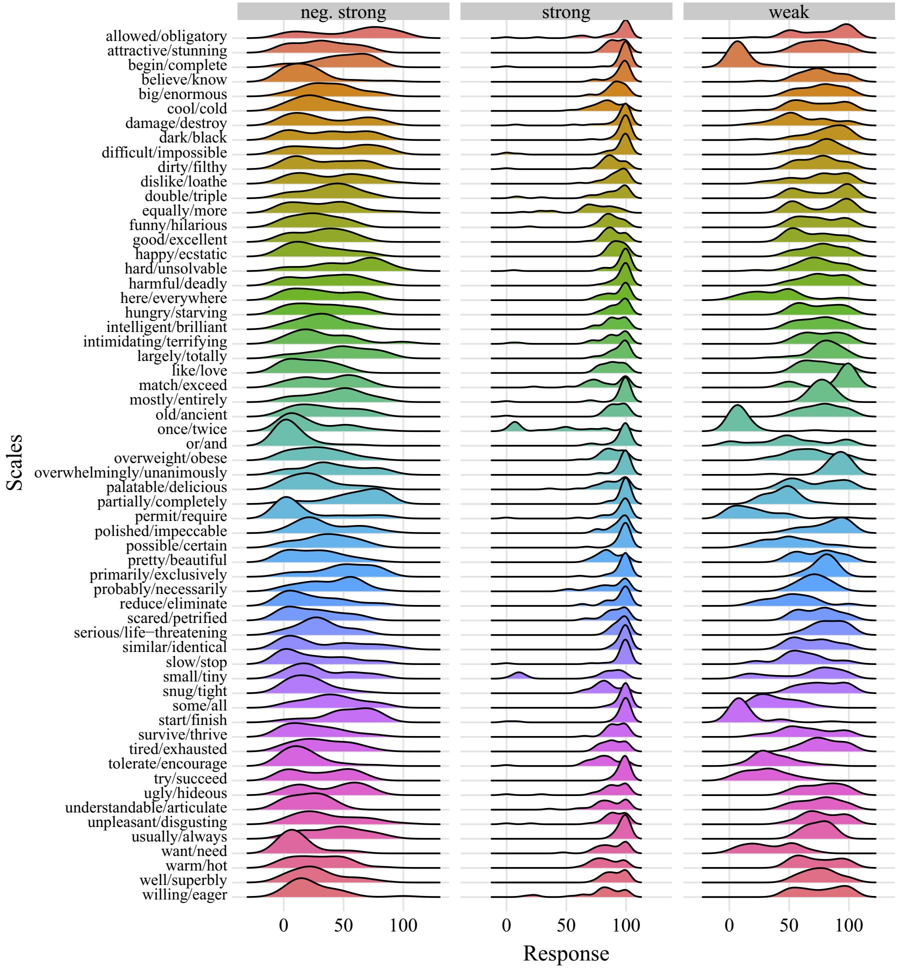


Figure A1. Distribution of Responses by-scale in Experiment 1.

https://doi.org/10.1017/langcog.2024.55 Published online by Cambridge University Press

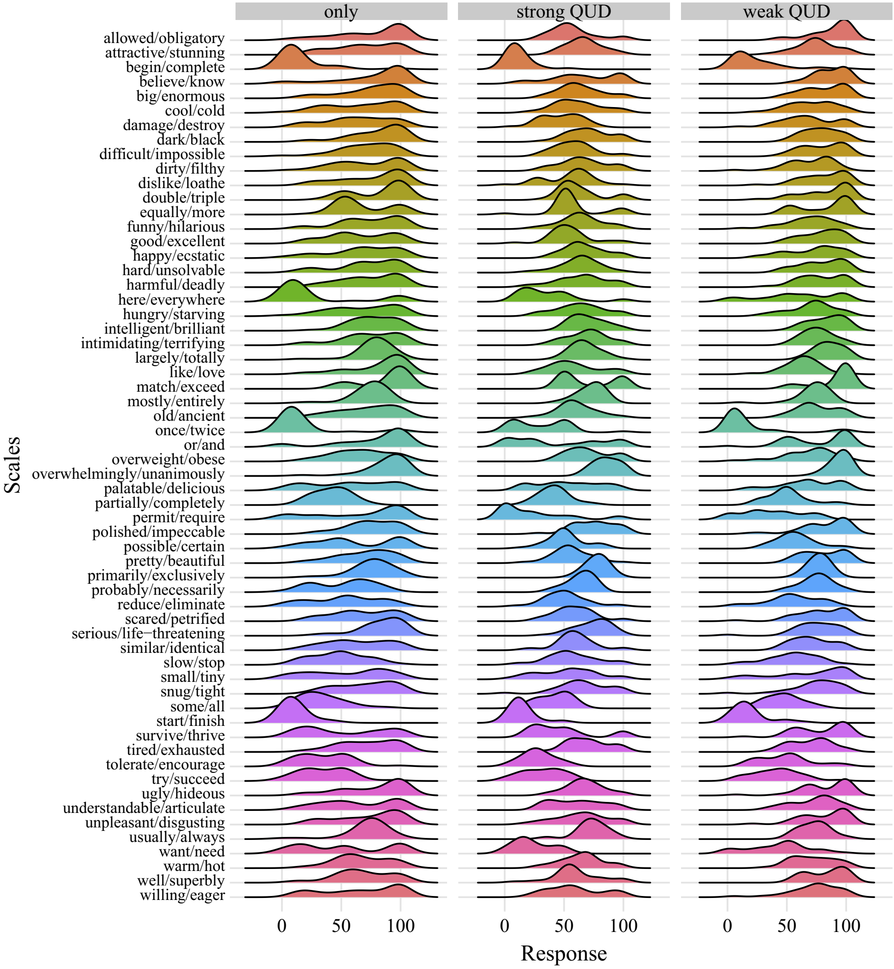


Figure A2. Distribution of Responses by-scale in Experiment 2.

Cite this article: Ronai, E., & Xiang, M. (2025). Scalar inference calculation through the lens of degree estimates, *Language and Cognition*, 17, e17, 1–26. https://doi.org/10.1017/langcog.2024.55