2022. *Proc Ling Soc Amer* 7(1). 5216. https://doi.org/10.3765/plsa.v7i1.5216.

Quantifying semantic and pragmatic effects on scalar diversity

Eszter Ronai & Ming Xiang*

Abstract. Scalar inference (SI), the process by which we systematically infer meanings stronger than what was explicitly said, has long been a central topic of investigation in semantics-pragmatics. A recent experimental finding that has generated interest is *scalar diversity*: that the robustness of SI calculation varies across lexical scales. For instance, the *some but not all* SI is much more likely to arise than *good but not excellent*. In this paper, we take a first step toward more rigorously quantifying the observed variation across scales using relative entropy. We then turn to the question of how factors independent of scalar diversity can make SI calculation both more likely and more uniform. We find that a supportive discourse context and overt exhaustification with the focus particle *only* both increase inference rates and reduce variation across scales, with the effect of *only* being stronger. However, there still remains a lot of scalar diversity; only when we combine context with semantic exhaustification do we find uniformity across lexical scales.

Keywords. experimental pragmatics; scalar inference; scalar diversity; Question Under Discussion; focus semantics

1. Introduction. Interpreting natural language meanings often involves not only interpreting what was said, but also what was left unsaid. A classic example of this kind of phenomenon is scalar inference (SI), exemplified in (1).

(1) Mary ate some of the cookies.
  a. Mary ate some, and possibly all, of the cookies. literal
  b. Mary ate some, but not all, of the cookies. SI

The literal meaning of an utterance like (1) is (1-a). But such an utterance also brings to mind an alternative that could have been said: *Mary ate all of the cookies*. This serves as an alternative because *some* and *all* form a scale: *all* is informationally stronger, and therefore more informative, than *some*. Hearers assume that speakers are trying to be maximally informative (following the Maxim of Quantity); therefore, if the alternative *Mary ate all of the cookies* were true, the speaker would have said that. Because she did not say it, hearers can infer its negation (following the Maxim of Quality). This reasoning process, combined with the utterance's literal meaning, leads to the SI-enriched meaning in (1-b) (Grice 1967; Horn 1972).

In experimental investigations of SI, a growing amount of attention is being paid to other pairs of lexical items that form a scale and lead to SI. One such example, based on the <*good*, *excellent*> scale, is given in (2).

(2) The movie is good.
  a. The movie is good, and possibly excellent. literal
  b. The movie is good, but not excellent. SI

* We would like to thank Zsolt Veraszto, Chris Kennedy, Michael Tabatowski, and the audiences at LSA 96, AMLaP 27, and at the Cornell University LiMe and C.Psyd labs for helpful discussion and feedback. 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, The University of Chicago (ronai@uchicago.edu) & Ming Xiang, The University of Chicago (mxiang@uchicago.edu).

Similarly to (1), (2) can also lead hearers to go beyond its literal meaning (2-a), via the same process of reasoning about an informationally stronger alternative. Specifically, hearers may reason about the stronger alternative *The movie is excellent*; because the speaker chose not to say this alternative, hearers can conclude that she must believe it to be false—leading to the SI in (2-b). But an influential experimental result from recent literature has revealed that lexical scales such as <*some*, *all*> vs. <*good*, *excellent*> differ substantially in how likely they are to lead to SI calculation (van Tiel et al. 2016; see also Baker et al. 2009; Doran et al. 2012; Beltrama & Xiang 2013).

The finding of scalar diversity has given rise to a research program of identifying factors that predict how likely a scale is to lead to SI calculation. Factors that have been put forward to explain scalar diversity make reference to properties of scales such as distinctness (van Tiel et al. 2016), accessibility (Ronai & Xiang to appear) or, focusing on adjectival scales, polarity and extremeness (Gotzner et al. 2018); to other semantic-pragmatic processes such as negative strengthening (Gotzner et al. 2018) or propensity for local enrichment (Sun et al. 2018); or to properties of the context (Pankratz & van Tiel 2021; Ronai & Xiang 2021a).

Our goal in this paper is different: rather than investigate what factors or properties of scales can predict scalar diversity, we look at how two factors unrelated to scalar diversity affect the likelihood and uniformity of calculating upper-bounded inferences such as *some but not all* or *good but not excellent*. We first manipulate the discourse context via an explicit question containing the stronger alternative (*all*, *excellent*), finding that such a supportive context both makes inference calculation more likely and reduces variability across scales. We then find similar, but more pronounced, effects by introducing overt exhaustification using the focus particle *only*: *only good*, *only some*. Lastly, we show that when these two factors align, i.e. there is pragmatic support from the context and semantic support from *only*, upper-bounded inferences are calculated at ceiling rates, and scalar diversity is eliminated. Crucially, our observations about whether variability is reduced are grounded in our proposal to more rigorously quantify scalar diversity than has been done in prior work.

This paper is structured as follows. In Section 2 we replicate scalar diversity (Experiment 1) and describe our proposal to use relative entropy to quantify it. In Section 3 we conduct a pragmatic context manipulation (Experiment 2), while in Section 4 we introduce a semantic manipulation using *only* (Experiment 3). In Section 5, we combine these manipulations (Experiment 4). Section 6 concludes.

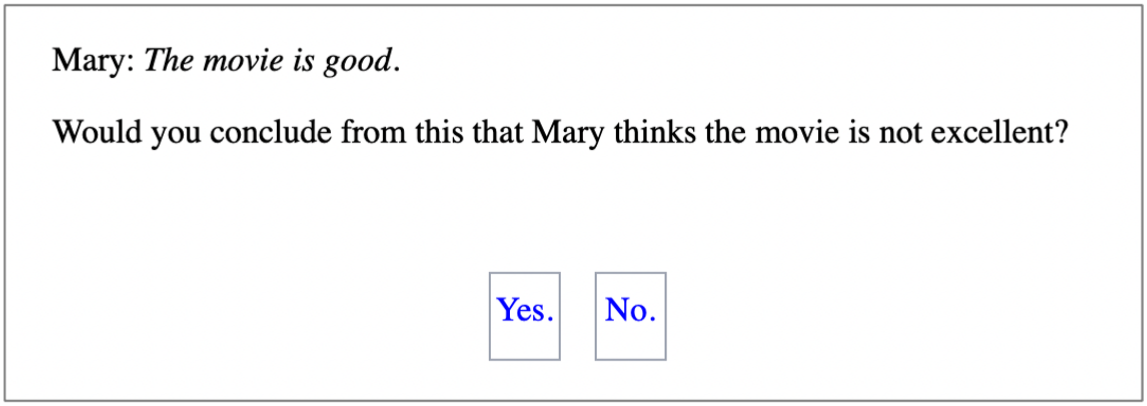
2. Experiment 1: Scalar diversity.

Experiment 1 was a replication of the basic scalar diversity effect, adapting the inference task used by van Tiel et al. (2016).

2.1. Participants and task.

42 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 40 participants is reported below.

An inference task was used to measure the rate of SI calculation. Participants were presented with trials such as Figure 1, where a speaker (Mary) uttered a potentially SI-triggering sentence like *The movie is good.* They then had to answer whether they would conclude that Mary thinks the movie is not excellent. In this two-alternative forced choice task, a "Yes" response indicates that the participant has calculated the SI: they are reasoning with the enriched *good but not excel-*



Mary: *The movie is good.*

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

Figure 1. Example experimental trial from Experiment 1

lent meaning of *good*, given Mary's utterance. A response of "No", on the other hand, indicates that they have not calculated the SI; the movie being good is compatible with the movie being excellent (according to Mary). The percentage of "Yes" responses is therefore our measure of the robustness of SI calculation.

The experiment tested 60 different lexical scales, which included some of the scales tested i.a. by van Tiel et al. (2016) and de Marneffe & Tonhauser (2019), as well as a selection found by querying the Corpus of Contemporary American English (Davies 2008)—for details of how our scale set was constructed, see Ronai & Xiang (to appear). In addition to the 60 critical items, 7 filler items were also included. These items served as catch trials and were meant to unambiguously elicit either a "Yes" or a "No" response, because they contained two terms either in an entailment relation (*wide* → *not narrow*) or unrelated to each other (*sleepy* → *not rich*). At the start of the experiment, participants saw 2 practice trials, which were followed by a total of 67 (critical and filler) trials.

2.2. Hypothesis and predictions. As reviewed in the Introduction, experimental studies have consistently found diversity in SI calculation across different lexical scales. We therefore expect to see variation across our 60 scales in how likely they are to lead to SI calculation, that is, in the percentage of "Yes" responses from the inference task.

2.3. Results and discussion. Figure 2 shows the results of Experiment 1 (leftmost facet: "Scalar inference"). As can be seen in the visualization, the rate of SI calculation, as indexed by the percentage of "Yes" responses, varies substantially across the 60 scales we tested. Specifically, SI rates range from 2.5% (<*scared*, *petrified*>; <*tired*, *exhausted*>) to 95% (<*partially*, *completely*>). This constitutes a replication of earlier findings of scalar diversity, by i.a. van Tiel et al. (2016).

Crucially, while the observation of scalar diversity was based only on descriptive statistics in previous work (range of SI rates), in this paper we propose a more rigorous measure to quantify the variation across scales. Specifically, we turn to information theoretic measures, which are commonly used, for instance, in the domain of syntactic processing (e.g., surprisal: Levy 2008; entropy reduction: Hale 2003). In particular, we propose using relative entropy (Kullback & Leibler 1951), a measure that compares two probability distributions and quantifies their difference. To quantify scalar diversity in Experiment 1, we treated the normalized SI rates (i.e.,

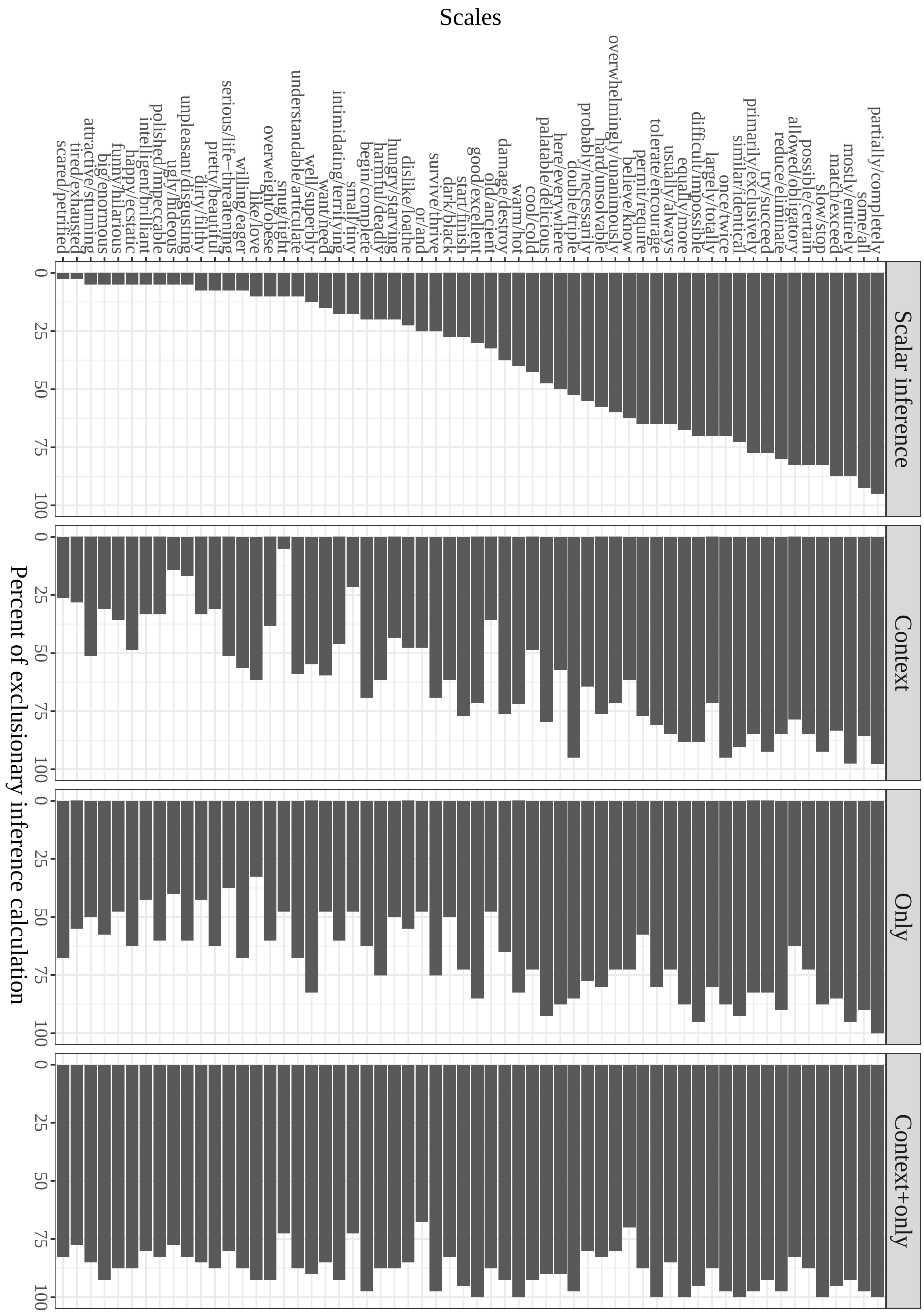


Figure 2. SI rate for 60 different scales. Experiments 1-4 are shown on the four facets of the plot.

percentage of "Yes" responses) across different scales as a probability distribution. We then compared this distribution to the uniform distribution. The uniform distribution represents a (hypothetical) scenario where each scale leads to the same SI rate. This scenario reflects the implicit assumption made by theoretical accounts, which suggest that SI calculation, which proceeds via reasoning about a stronger scalar alternative, should not vary across different scales—the so-called "uniformity assumption" (van Tiel et al. 2016; p. 139). As we quantify diversity via comparison to a uniform distribution, we do not assume any particular SI rate as the basis for uniformity; in our calculation of relative entropy, we remain agnostic about whether "uniform" would mean 100% SI rate across all scales, 70%, or so on. Intuitively, relative entropy represents how "surprised" we are if we assume a particular distribution (the uniform distribution), but observe a different one (Experiment 1).

The equation in (3) was used to calculate relative entropy. Here, $p(x)$ is the normalized observed percentage of "Yes" responses across scales in Experiment 1, $\mathcal{X}$ is the 60 scales, i.e., the finite set over which we defined our probability distribution, and $q(x) = 1/60$ is the uniform probability mass function over the 60 scales. In this specific case, because the uniform inference rate is a constant across all 60 scales, the relative entropy that we obtain is the entropy of the uniform distribution minus the entropy of the experimentally collected SI rates.

(3) Let $p(x)$ and $q(x)$ be probability mass functions over the same set $\mathcal{X}$. The relative entropy of $p(x)$ with respect to $q(x)$ is given by

$$D\left(p||q\right) = \sum_{x \in \mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right).$$

The relative entropy of our Experiment 1 SI rates is 0.466. To contextualize this number, we may consider a number of hypothetical scenarios as benchmarks. If all scales indeed led to the same rate of SI calculation, then that would give a relative entropy of 0—see the Benchmark 1 facet in Figure 3. On the other extreme, the highest possible relative entropy would be obtained if all the probability mass was concentrated on a single scale: that is, if only one of the 60 scales ever led to SI calculation (at some non-zero rate), while the other 59 scales did not—this hypothetical scenario would lead to a relative entropy of 5.907, and it is shown as Benchmark 2 in Figure 3. Closer to the actual experimental findings is Benchmark 3: a hypothetical "linear" distribution, where likelihood of SI calculation is evenly distributed across the 60 lexical scales over a 0-100 scale. Here, for instance, one scale leads to SI calculation at a 1.67% rate, the next at 3.33%, the one after that at 5%, etc., up to scale number 60 leading to SI calculation at 100%. This linear benchmark would yield a relative entropy of 0.2912. Lastly, the "quadratic" distribution in Benchmark 4 is a scenario similar to Benchmark 3, in that every scale has a unique SI calculation rate; but here, probability mass is more concentrated toward one scale, giving a relative entropy of 0.6352. We can see that the experimentally collected rates fall between Benchmarks 3 and 4 (0.466), suggesting more diversity than Benchmark 3, but less than 4.

The benchmarks outlined here are for illustration; the main goal of this paper is to use the proposed relative entropy measure to compare different sets of experimentally collected SI rates to one another, seeing how different manipulations reduce variation across lexical scales. This is what we now turn to in Experiment 2.

3. Experiment 2: Discourse context manipulation. Much of the experimental pragmatics literature, including on scalar diversity, and including our Experiment 1, presents stimulus sentences

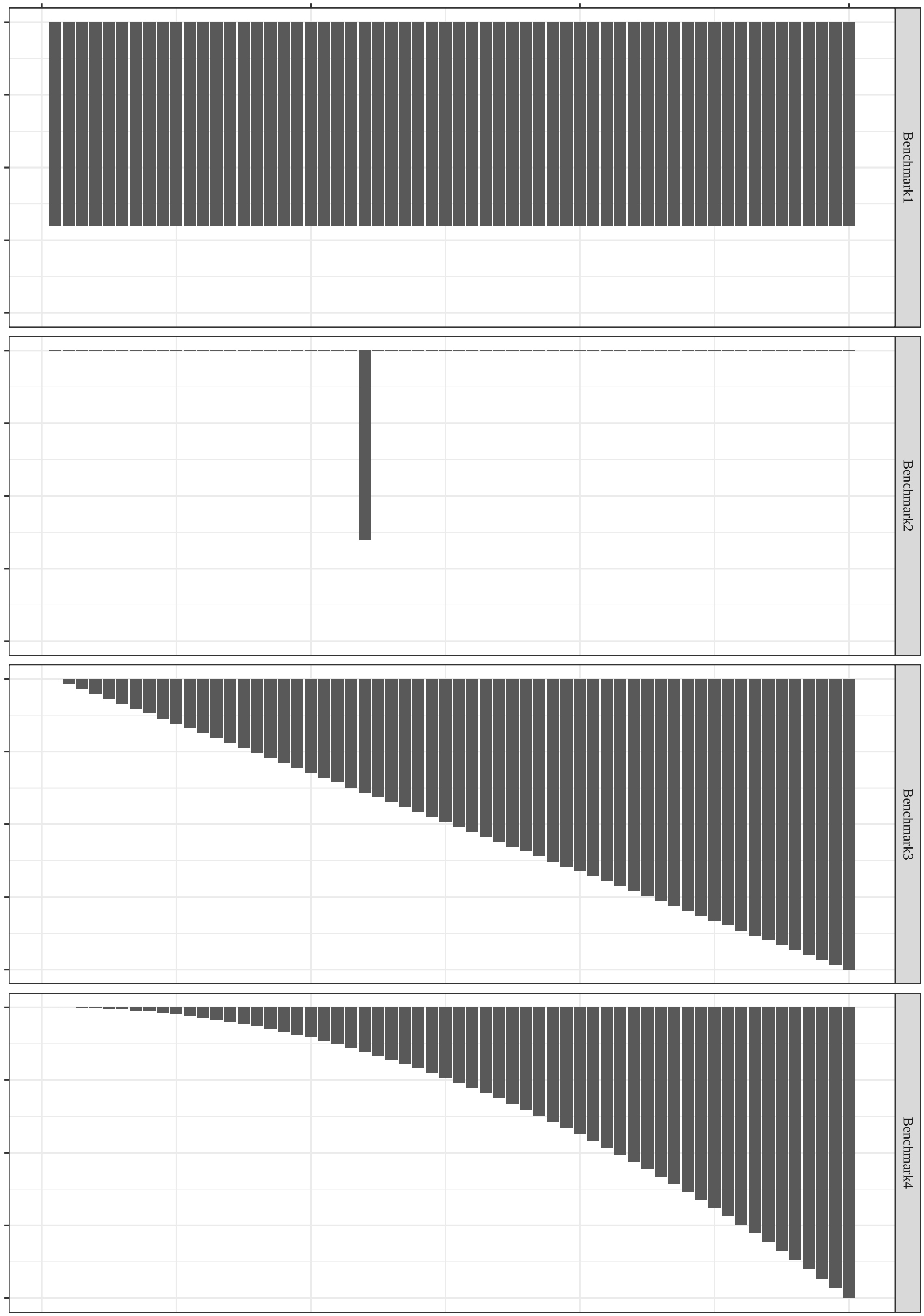


Figure 3. Hypothetical SI rates: benchmarks for quantifying diversity using relative entropy.

in the absence of any context. But it is well known that properties of the discourse context, formalized e.g. as Question Under Discussion (Roberts 1996/2012), can make SI calculation more or less likely (Kuppevelt 1996). Indeed, experimental work has confirmed this modulating role of context not only for the <*some*, *all*> scale (Degen & Tanenhaus 2015; Ronai & Xiang 2021b; Yang et al. 2018; Zondervan et al. 2008), but also for a variety of different lexical scales (Cummins & Rohde 2015; Ronai & Xiang 2021a).

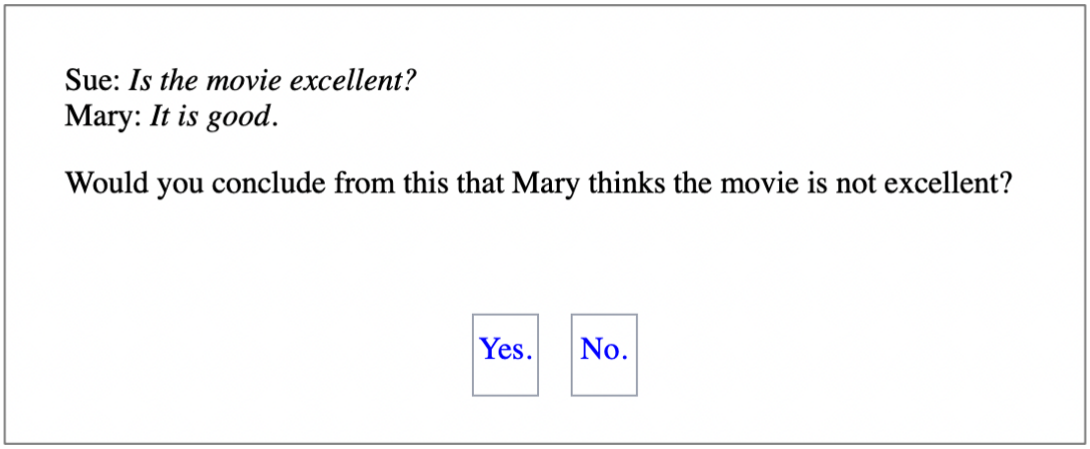
In Experiment 2, we operationalize discourse context as explicit questions, and investigate the effect of such a manipulation on the likelihood of SI calculation, as well as on the observed variation across scales, that is, scalar diversity.

3.1. PARTICIPANTS AND TASK. 81 native speakers of American English participated in an online experiment, administered on the Ibex (Drummond 2007) and PCIbex (Zehr & Schwarz 2018) platforms. Participant recruitment, screening, and compensation was identical to Experiment 1. Data from all 81 participants is reported below.

Experiment 2 employed the same task as Experiment 1, but the potentially SI-triggering sentences (uttered by Mary) were now embedded in a dialogue context. Specifically, the SI-triggering sentences were preceded by a polar question that contained the stronger scalar term. For the <*good*, *excellent*> scale, for instance, the manipulation included the question *Is the movie excellent?*—see Figure 4. (Experiment 2 also included a within-participants condition where the explicit question contained the weaker scalar term, i.e., *Is the movie good?*, but in this paper we set those data aside and concentrate on the manipulation exemplified in Figure 4.)

Mary's answers were modified to ensure dialogue coherence, e.g., *The movie is good* was changed to *It is good*. Otherwise, Experiment 2's materials and procedure were identical to Experiment 1.

3.2. HYPOTHESIS AND PREDICTIONS. Standard semantic treatments of questions take them to partition a set of possible worlds into cells denoting their possible answers (Hamblin 1976; Groenendijk & Stokhof 1984). The question *Is the movie excellent?*, then, partitions the Common Ground based on the stronger alternative *excellent*: in one cell are all the worlds where the movie is excellent, and in the other cell, all the worlds where the movie is not excellent. An answer, in turn, is taken to be congruent with (or "a good answer to") a question if it determines which cell



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

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

Figure 4. Example experimental trial from Experiment 2

contains the actual world (Hulsey et al. 2004). Consider the two readings (literal and SI-enriched) of a potentially SI-triggering sentence in this light:

(4) The movie is good.
  a. The movie is good, and possibly excellent. literal
  b. The movie is good, but not excellent. SI

Given the question *Is the movie excellent?*, the SI-enriched meaning in (4-b) is a congruent answer, because it entails the "not excellent" cell of the partition, and eliminates the "excellent" cell. The literal meaning in (4-a), on the other hand, does not entail either cell, and it therefore does not directly bear on the question. Therefore, only on its SI-enriched meaning does *The movie is good* constitute a congruent answer.

Given this, we hypothesize that the discourse context manipulation in Experiment 2 will encourage SI calculation; that is, participants will calculate SIs in order to make Mary's answers congruent with Sue's questions. Consequently, we first predict that SI rates will increase as compared to the baseline Experiment 1, which included no context. In fact, inference rates could increase to ceiling (100%), since without SI calculation, the dialogue participants are presented with would not be congruent. Second, we predict that as inference rates increase across the board for all scales, variation across scales (scalar diversity) will be reduced.

3.3. RESULTS AND DISCUSSION. Figure 2 shows the results of Experiment 2 (second facet: "Context"). To compare the rates of inference calculation in Experiment 2 to the rates from Experiment 1, we fit a logistic mixed effects regression model using the lme4 package in R (Bates et al. 2015). The model predicted Response ("Yes" vs. "No") as a function of Experiment. We included the maximal random effects structure supported by the data (Barr et al. 2013): random intercepts for participants and random slopes and intercepts for items. The fixed effects predictor Experiment (1 vs. 2) was sum-coded before analysis, with Experiment 1 mapping to negative and Experiment 2 to positive coefficients. This analysis revealed an overall increase in inferences rates in Experiment 2, as compared with Experiment 1 (Estimate=1.49, SE=0.22, z=6.56, p <0.001). That is, participants were significantly more likely to calculate inferences in a supportive discourse context. This finding replicates Ronai & Xiang's (2021a) Experiment 2 on a different, larger set of scales.

To check the effect of the discourse context manipulation on the variation in SI rates across scales, we can turn to our measure of relative entropy. The SI rates in Experiment 2 resulted in a relative entropy of 0.126. Recall that lower numbers represent more uniformity: if all scales led to SI at the same rate, relative entropy would be 0, but the relative entropy of the baseline Experiment 1 (without context) was 0.466. What we find, then, is that the context manipulation reduced the variation in SI rates: relative entropy is lower in Experiment 2 than in 1, i.e., there is less scalar diversity. Altogether, in line with our predictions, an explicit question based on the stronger scalar term both increased SI rates across the board and reduced the variation across scales.

At the same time, however, we did not find a ceiling effect in SI rates, nor did we find uniformity across scales. This presents a puzzle. As discussed, a question such as *Is the movie excellent?* partitions the Common Ground into possible worlds where the movie is excellent vs. possible worlds where the movie is not excellent. Given this, Mary's utterance only constitutes a felicitous contribution to the discourse on its SI-enriched meaning (*good but not excellent*), because only on this meaning does it entail one of the cells of the partition. Since this is the same for all scales tested, we would expect equally high SI calculation everywhere. We propose the following possible reason for why the predicted uniformity does not obtain: there are in fact three different possible pragmatic meanings that can be attributed to Mary's utterance in the dialogue context, which we detail below (5-a)-(5-c).

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

a. It is good (but not excellent). SI
b. (Well,) it's good. ignorance
c. (Yes,) it's good. good ≈ excellent

(5-a) is the standard scalar implicature meaning, which was the one we intended for participants to arrive at in Experiment 2. It is also possible, though, that (some) participants assigned to Mary's answer the meaning in (5-b), which is communicating ignorance about the stronger alternative; on this reading, Mary's answer conveys not that the movie is not excellent, but that Mary does not know whether it is excellent. Lastly, (5-c) shows a third possibility, where *good* is used as a synonym for *excellent*—Mary is in fact giving an affirmative answer to Sue's question. Using a weaker scalar term as a synonym for a stronger alternative may be related to the semantic distance between or distinctness of the two scalar terms, which has been shown to independently correlate with SI rates: the more distant or distinct the two terms, the more likely the SI (van Tiel et al. 2016; Ronai & Xiang to appear). It is also possible to analyze the interpretation in (5-c) as an R-implicature (Horn 1984). (6) shows a classic example of an R-implicature, where a statement of (6-a) implicates (6-b).

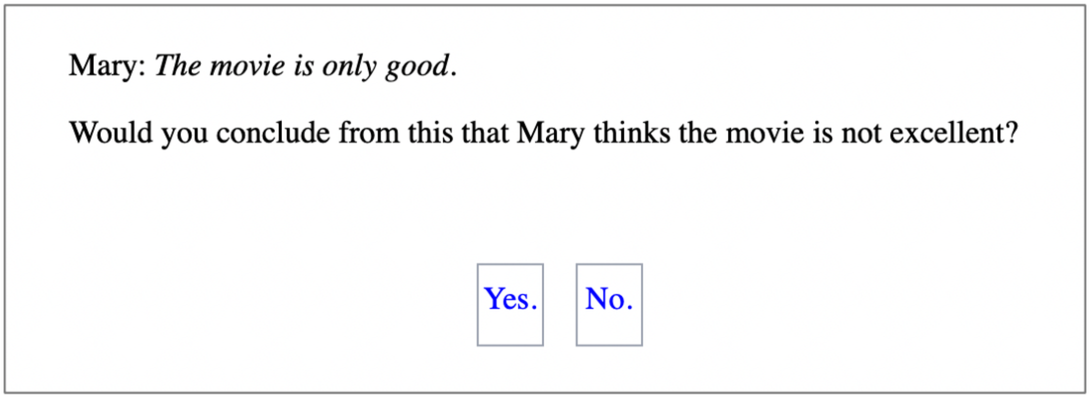
(6) a. I need a drink.
b. I need an alcoholic drink.

In R-implicature, a generic form (*drink*) takes on a more specific meaning (*alcoholic drink*). In other words, while scalar implicatures introduce upper-bounded meanings (*good* means *not more than good*), R-implicatures are lower-bounding: what is said is the lower limit of what is actually the case. Arguably, taking *The movie is good* to affirm that the movie is excellent is an instance of such an implicature.

Even though our experimental manipulation intended for participants to arrive at the meaning in (5-a), we must note that (5-a) and (5-c) both represent congruent answers: (5-a) addresses the question by entailing "not excellent", while (5-c) addresses it by entailing "excellent". But only if a participant had (5-a) in mind did they answer "Yes"; with either (5-b) or (5-c), they answered "No". Crucially, all three different readings for Mary's *It is good* answer should correspond to different prosodic contours; directly manipulating prosody using audio stimuli to tease them apart is therefore a promising area for future work.

4. Experiment 3: Focus manipulation. While Experiment 2 tested a pragmatic manipulation, Experiment 3 uses a semantic one: we investigate the effect of focus (encoded by the particle *only*) on the likelihood and diversity of inference calculation.

4.1. PARTICIPANTS AND TASK. 41 native speakers of American English participated in an online (Ibex) experiment for $2 compensation. Participant recruitment, screening, and compensation was identical to Experiment 1. Data from 40 participants is reported below.



Mary: *The movie is only good.*

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

Figure 5. Example experimental trial from Experiment 3

Experiment 3 employed the same inference task as the previous two experiments. This time, the additional manipulation we conducted was to include the focus particle *only* in the SI-triggering statement. That is, Mary's utterance was e.g., *The movie is only excellent*—see Figure 5 for an example trial. Other than this, the materials and procedure were identical to Experiment 1.

4.2. Hypothesis and predictions. The focus operator *only* semantically excludes alternatives to the focused element (Rooth 1985, 1992). Unlike in previous experiments, where the exclusion of the stronger alternative (e.g., *excellent*) was a cancellable pragmatic inference, in Experiment 3, alternative exclusion is an entailment. Based on this, we predict that comprehenders will exclude alternatives to the focused element, and consequently that the rates of upper-bounded inference calculation will increase. Again, inference rates could possibly increase to ceiling (100%), given that *The movie is only good* encodes the exclusion of alternatives to *good* in a non-cancellable way. As a consequence of inference rates increasing across the board, we also predict that the variation (diversity) observed across lexical scales will be reduced.

4.3. Results and discussion. Figure 2 shows the results of Experiment 3 (third facet: "Only"). To compare Experiment 3's results to that of the baseline Experiment 1, we conducted the same statistical analysis as the one reported in Section 3.3. This analysis revealed that Experiment 3 also led to significantly higher rates of inference calculation than Experiment 1 (Estimate=1.86, SE=0.27, z=6.96, p <0.001)—participants were more likely to calculate upper-bounded inferences in the presence of overt exhaustification with *only*. Turning now to our measure of the "diversity" of inference rates, Experiment 3 led to a relative entropy value of 0.046. Compared to the previous experiments (see Table 1), we see a more substantial reduction in variation across scales; scalar diversity was lessened more with the focus manipulation than with the discourse context manipulation.

	Manipulation	Relative entropy
Experiment 1	Baseline scalar diversity	0.466
Experiment 2	Discourse context	0.126
Experiment 3	Exhaustification with *only*	0.046
Experiment 4	Context and *only*	0.006

Table 1. Relative entropy results by experiment

In line with our predictions, then, the focus manipulation made upper-bounded inference calculation (*some but not all*, *good but not excellent*) more likely, and it also reduced scalar diversity. However, as is evident from Figure 2, we do not have ceiling-level inference rates: it is not the case that encoding alternative exclusion in the semantics always led participants to answer "Yes" in the inference task for all scales. Additionally, there still remains variability across the different scales. We propose two potential (related) explanations for this. First, *only* is ambiguous between its so-called rank-order reading (i.a. Horn 2000) and its complement-exclusion reading (i.a. Hole 2004). The rank-order reading of *only* can be paraphrased as *no more than*: *The movie is only good* means that the movie is no more than good. On this reading, the stronger alternative *excellent* must be excluded; if participants were all assigning this meaning to *only*, we should be seeing ceiling level "Yes" responses. On the other hand, however, the complement-exclusion reading of *only* has the meaning *nothing other than*: *The movie is only good* means that the movie is nothing other than good. On this reading, it is possible to interpret Mary's utterance in our experiment as communicating that the movie is good, but not funny or thrilling, etc. Excluding such non-scalar alternatives leaves open the possibility that the movie is in fact excellent, leading participants to respond "No" in the inference task.

Relatedly, in Experiment 3, stimulus sentences appeared without any context. It is therefore possible that participants had different contexts in mind. Compare, for instance, (7) and (8).

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

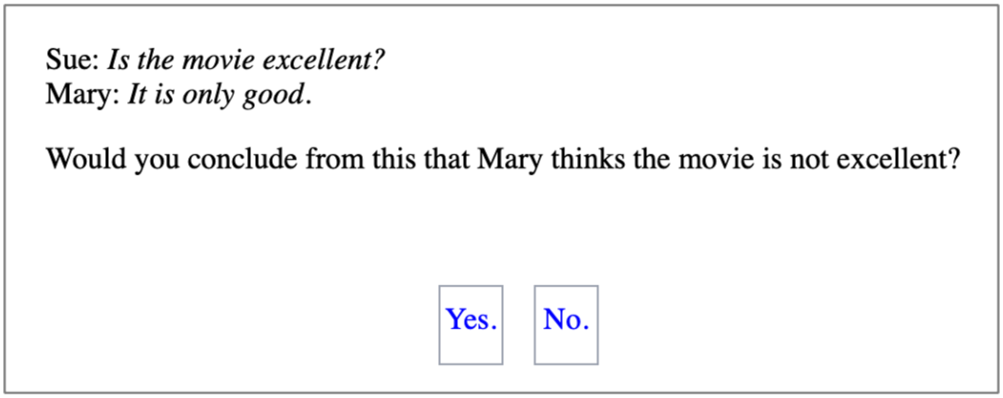
(8) Sue: What's the movie like?
Mary: It is only good.

A context like (7) makes salient *excellent* as an alternative. Given such a context, *only* is most naturally interpreted as excluding this alternative *excellent*. If a participant supposed such a context, then, they would arrive at the upper-bounded *good but not excellent* inference, and answer "Yes" in the experimental task. But if they supposed a context like the one in (8), the alternatives that are to be excluded could be any property that a movie can have. Therefore, on this interpretation, participants could have concluded that the movie is only good, but not funny, thrilling, scary, etc., ultimately answering "No" to the task question about the movie not being excellent.

To summarize, while overt exhaustification with the focus particle *only* encodes alternative exclusion in the semantics, it is possible that participants interpreted Mary's utterance as excluding some alternative(s) other than the one we tested (i.e., something other than *excellent* for *good*). We argue that this might be the reason inferences rate were not at ceiling in Experiment 3, and why some of the inter-scale variation still remained[1].

5. Experiment 4: Manipulating both cues. Experiments 2-3 showed that a pragmatic manipulation (supportive discourse context) and a semantic one (exhaustification with *only*) both increase inference rates and reduce variation across lexical scales. However, we also saw that with either manipulation, some of the variation still remains. Experiment 4 therefore combines context with *only*.

[1] For a small minority of our items, it is also possible that there was ambiguity not in the identity of the alternative, but in the focus associate itself. For example, the sentence *The princess only likes dancing* (intended SI: *She doesn't love dancing*) could also be interpreted such that *only* associates not with *like*, but with *dancing*, leading to the inference that the princess does not like activities other than dancing.



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

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

Figure 6. Example experimental trial from Experiment 4

5.1. Participants and task. 40 native speakers of American English participated in an online (Ibex and PCIbex) experiment for $2 compensation. Participant recruitment, screening, and compensation was identical to Experiment 1. Data from all 40 participants is reported below.

Experiment 4 combined the manipulation of Experiments 2 and 3: the potentially inference-triggering sentences included the focus particle *only*, and they were also preceded by a polar question that made reference to the stronger scalar alternative—see Figure 6 for an example. Otherwise, the task and materials were identical to previous experiments.

5.2. Hypothesis and predictions. The predictions made for Experiments 2 and 3 straightforwardly carry over to Experiment 4. First, because of Sue's explicit question, only on its inference-enriched meaning is Mary's answer congruent, which predicts increased inference rates and less variation across scales (see Section 3.2 for details). Second, *only* encodes the exclusion of alternatives in the semantics, which similarly predicts increased inference rates and less variation (Section 4.2).

Moreover, in Section 4.3 we argued that the reason Experiment 3's manipulation with *only* did not lead to uniformity and ceiling-level inference rates is that participants may have been excluding non-scalar alternatives. The discourse manipulation of Experiment 4 makes clear the identity of the intended alternative, and we therefore predict that the variation that remained in Experiment 3 should be eliminated.

5.3. Results and discussion. Figure 2 shows the results of Experiment 4 (rightmost facet: "Context + only"). A statistical analysis identical to the one reported in Section 3.3 confirms that Experiment 4's manipulation significantly increased rates of inference calculation as compared to Experiment 1's baseline (Estimate=3.74, SE= 0.35, z=10.64, p <0.001). As can be seen in the figure, inference rates are now in fact almost at ceiling.

The relative entropy resulting from Experiment 4 is 0.006—see Table 1 for a comparison of the relative entropy from all experiments. We can see that uniformity was very nearly achieved in Experiment 4; we no longer find appreciable variation in inference rates across the lexical scales tested.

Overall, Experiments 2–4 are informative as to the interplay of different factors that can make the calculation of upper-bounded inferences more likely and more uniform. We have seen that discourse context can provide salient alternatives; but it does not tell hearers that they need to reason about and exclude them. The focus particle *only*, on the other hand, makes reasoning about and excluding alternatives obligatory, but it does not make it clear what the relevant alternatives are. As Experiment 4 demonstrates, only when both of these factors are fixed – the identity of the alternatives is made clear, and the cue to exclude them is encoded semantically – do we find ceiling effects in inference calculation. When either of these supportive cues is absent, there is more flexibility in interpretation, and consequently we observe more variation. This also leaves more opportunity for other factors (reviewed in Section 1, e.g., distinctness, extremeness, etc.) to influence the likelihood of inference calculation.

6. Conclusion. Previous research has revealed that different scalar expressions give rise to SI at different rates; however, this observation of scalar diversity has so far been based on descriptive statistics. In this paper, we offered a more rigorous way to quantify the diversity of SI rates using relative entropy. Additionally, we investigated what factors can increase SI rates and introduce uniformity. Our findings revealed that a pragmatic manipulation (explicit question) and a semantic manipulation (exhaustification with *only*) both lead to increased inference rates and reduced diversity—the latter more so than the former. However, variation still remains under either manipulation, and only when we combine them do we find ceiling level inference rates and uniformity across scales.

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.

Barr, Dale J., Roger Levy, Christoph Scheepers & Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3). 255–278. https://doi.org/10.1016/j.jml.2012.11.001.

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.

Cummins, Chris & Hannah Rohde. 2015. Evoking context with contrastive stress: Effects on pragmatic enrichment. *Frontiers in Psychology* 6. 1779. https://doi.org/10.3389/fpsyg.2015.01779.

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

Degen, Judith & Michael K. Tanenhaus. 2015. Processing scalar implicature: A constraint-based approach. *Cognitive Science* 39(4). 667–710. https://doi.org/10.1111/cogs.12171.

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.

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.

Groenendijk, Jeroen & Martin Stokhof. 1984. On the semantics of questions and the pragmatics of answers. In Fred Landman & Frank Veltman (eds.), *Varieties of formal semantics: Proceedings of the Fourth Amsterdam Colloquium*, 143–170. Dordrecht: Foris.

Hale, John. 2003. The information conveyed by words in sentences. *Journal of Psycholinguistic Research* 32(2). 101–123. https://doi.org/10.1023/A:1022492123056.

Hamblin, Charles L. 1976. Questions in Montague English. In Barbara H. Partee (ed.), *Montague grammar*, 247–259. New York: Academic Press.

Hole, Daniel. 2004. *Focus and background marking in Mandarin Chinese: System and theory behind* cai, jiu, dou *and* ye. London: Routledge. https://doi.org/10.4324/9780203565193.

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

Horn, Laurence R. 1984. Toward a new taxonomy for pragmatic inference: Q-based and r-based implicature. In Deborah Schiffrin (ed.), *Meaning, form, and use in context: Linguistic applications*, 11–42. Washington D.C.: Georgetown University Press.

Horn, Laurence R. 2000. Pick a theory (not just any theory): Indiscriminatives and the free-choice indefinite. In Laurence R. Horn & Yasuhiko Kato (eds.), *Negation and polarity: Syntactic and semantic perspectives*, 147–192. Oxford, UK: Oxford University 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.

Kullback, Solomon & Richard A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22(1). 79–86.

Kuppevelt, Jan van. 1996. Inferring from topics: Scalar implicatures as topic-dependent inferences. *Linguistics and Philosophy* 19(4). 393–443.

Levy, Roger. 2008. Expectation-based syntactic comprehension. *Cognition* 106(3). 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006.

de Marneffe, Marie-Catherine & Judith Tonhauser. 2019. Inferring meaning from indirect answers to polar questions: The contribution of the rise-fall-rise contour. In Malte Zimmermann, Klaus von Heusinger & Edgar Onea (eds.), *Current research in the semantics/pragmatics interface* (Questions in Discourse 36), 132–163. Leiden, The Netherlands: Brill. https://doi.org/10.1163/9789004378322 006. _

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. 2021a. Exploring the connection between Question Under Discussion and scalar diversity. *Proceedings of the Linguistic Society of America* 6(1). 649–662. https://doi.org/10.3765/plsa.v6i1.5001.

Ronai, Eszter & Ming Xiang. 2021b. Pragmatic inferences are QUD-sensitive: An experimental study. *Journal of Linguistics* 57(4). 841–870. https://doi.org/10.1017/S0022226720000389.

Ronai, Eszter & Ming Xiang. To appear. Three factors in explaining scalar diversity. In Daniel Gutzmann & Sophie Repp (eds.), *Proceedings of Sinn und Bedeutung 25*.

Rooth, Mats. 1985. Association with focus. Amherst: University of Massachusetts, Amherst dissertation.

Rooth, Mats. 1992. A theory of focus interpretation. *Natural Language Semantics* 1(1). 75–116. https://doi.org/10.1007/bf02342617.

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, Emiel Van Miltenburg, Natalia Zevakhina & Bart Geurts. 2016. Scalar diversity. *Journal of Semantics* 33(1). 137–175. https://doi.org/10.1093/jos/ffu017.

Yang, Xiao, Utako Minai & Robert Fiorentino. 2018. Context-sensitivity and individual differences in the derivation of scalar implicature. *Frontiers in Psychology* 9. 1720. https://doi.org/10.3389/fpsyg.2018.01720.

Zehr, Jeremy & Florian Schwarz. 2018. PennController for Internet Based Experiments (IBEX). https://doi.org/10.17605/OSF.IO/MD832.

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. https://doi.org/10.3765/salt.v18i0.2486.