What could have been said? Alternatives and variability in pragmatic inferences

Eszter Ronai & Ming Xiang

Abstract

A recent influential experimental finding in pragmatics is that of *scalar diversity*: that different lexical items vary robustly in how likely they are to lead to scalar inference. For instance, hearers are much more likely to strengthen the meaning of *some* to *some but not all* than to infer *good but not excellent* from *good*. In this paper, we address the question of what underlies scalar diversity and identify two sources of uncertainty: uncertainty associated with the identity of relevant alternatives, and uncertainty associated with the step of excluding those alternatives. In our experiments, we make use of the Question Under Discussion to eliminate the former, and of the focus particle *only* to eliminate the latter kind of uncertainty. Our findings show that both manipulations make inference calculation more likely, but only when they are combined is scalar diversity reduced to a minimum. In order to quantitatively characterize the observed (reduction in) variation, this paper adopts the information theoretic measure of relative entropy.

Keywords: pragmatics; scalar inference; scalar diversity; uncertainty; discourse context

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 at least some 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 than *some*, since a sentence like *Mary ate all of the cookies* entails *Mary ate some of the cookies*, but not vice versa (Horn, 1972). 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).

Going beyond the classic <*some*, *all*> example, a growing body of work looks at a larger range of lexical items that form a scale and potentially 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 at least good. literal

b. The movie is good, but not excellent. SI

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: hearers calculate the *some but not all* SI much more robustly

than the *good but not excellent* SI. In the first large-scale investigation of this inter-scale variation, van Tiel *et al.* (2016) tested 43 different lexical scales and found that the rate of SI calculation ranged from 4% to 100% (see also Baker *et al.* 2009; Doran *et al.* 2012; Beltrama & Xiang 2013; Simons & Warren 2018).

The finding of such robust variation is significant as it challenges the so-called uniformity assumption (van Tiel *et al.*, 2016, p. 139): a (tacit) assumption in prior literature that since all instances of SI are derived via the same mechanism —e.g., by hearers' reasoning that an informationally stronger unsaid alternative is not true —there should be no differences across scales. As van Tiel *et al.* (2016) have shown, however, instead of uniformity, we find *scalar diversity*. The finding of scalar diversity has since given rise to a research program of identifying parameters along which lexical scales differ, which can then predict how likely a scale is to lead to SI calculation, ultimately explaining inter-scale variation. Factors that have been put forward to explain scalar diversity make reference to properties of scales such as the distinctness of the weaker and stronger scale-mates (van Tiel *et al.*, 2016), their semantic relatedness (Westera & Boleda, 2020), the polarity of adjectival scales (Gotzner *et al.*, 2018), or whether the stronger alternative is an extreme adjective (Gotzner *et al.*, 2018; Beltrama & Xiang, 2013). Scalar diversity has also been related 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, broadly construed (Pankratz & van Tiel, 2021; Ronai & Xiang, 2021a). However, most of the predictors identified in previous studies still only explain a small amount of the "diversity". For example, van Tiel *et al.* found that the two components of distinctness, semantic distance and boundedness, accounted for 3% and 10% of the observed variance respectively.

One interesting hypothesis about scalar diversity focuses on the observation that there is substantial uncertainty regarding alternatives. To derive an SI from an utterance, hearers need to identify what is a relevant alternative to what was uttered, and then they also need to have good reasons to believe that the stronger alternative needs to be ruled out. Several authors have observed that it is not always obvious what the relevant unsaid alternatives are, whose negation can be inferred. Van Tiel et al. (2016) originally put this hypothesis in terms of alternative availability: for SI to arise, it has to be the case that the stronger alternative was available to the speaker, so she could have actually considered using it. While none of the empirical measures of alternative availability considered by van Tiel *et al.* turned out to be significant predictors of SI rates across scales, later work has provided supportive evidence for the role of alternative availability or uncertainty. In a recent study, Hu *et al.* (2023) (see also Hu *et al.* 2022) used large language models to test the hypothesis that hearers

maintain cue-based expectations over alternatives (Degen & Tanenhaus, 2015, 2016). They found variation in how expected an alternative is as a scale-mate, and that these differences can predict both intra-scale (Degen, 2015), and most importantly for the present paper, inter-scale variation in SI rates. Concretely, the authors show that the more expected the stronger scale-mate, given a weaker scalar term and the sentential context, the higher the SI rate. Similarly, Ronai & Xiang (2022) have provided experimental evidence for van Tiel *et al.*'s notion of alternative availability. The authors conducted a cloze task in a discourse context, where participants saw dialogues such as "A: *The movie is good.*"; "B: *So you mean it's not BLANK.*" and were asked to fill in the blank. The frequency with which the stronger alternative (here, *excellent*) was provided was taken to index the accessibility of alternatives, that is, how strongly the weaker scalar evokes the stronger alternative. Inter-scale variation in alternative accessibility was shown to predict inter-scale variation in SI rates. Moreover, Hu *et al.* (2023) found that this experimental data was significantly correlated with language model-based measures of alternative uncertainty, suggesting that "models and humans are aligned at the level of predictive distributions over alternatives" (p. 8).

Informative as they are, previous studies on how alternative uncertainty impacts inference calculation still leave open questions about the sources of uncertainty. One possibility we will examine in the current paper is that contextual relevance constrains the space of possible alternatives, and thus directly modulates the extent of scalar diversity. As we will discuss below, the effect of context is well-documented for SI calculation in general (Matsumoto, 1995; Van Kuppevelt, 1996; Degen, 2013, 2015). Pertaining to scalar diversity, there have also been suggestions that alternative uncertainty is contextually driven. McNally (2017), for instance, notes that due to polysemy, there might be context-based variation in what counts as a stronger alternative. While <*warm*, *hot*> form a scale based on asymmetric entailment, and *hot* is indeed a relevant stronger alternative to *warm* in the context of *The weather is warm*, this might not always be the case for *The soup is warm* (McNally, 2017, p. 23-24). This is because, as McNally argues, in the latter case, the choice of the adjective *warm* can be interpreted as referring to a kind of soup, in contrast to cold soups as a kind. Consequently, *The soup is warm* here means that it is a soup consumed warm (or possibly hot), not cold. Since *hot* is not necessarily a relevant alternative to *warm* in this case, the *not hot* SI may not arise. An informative study here is Pankratz & van Tiel (2021), which examined the contextual relevance of different SIs and found it to be a predictor of scalar diversity. The authors developed a corpus-based measure of relevance, whereby the more relevant an SI is, the more likely it is to occur in so-called scalar constructions, e.g., *good but not excellent*, *good rather than excellent*, or *good, if not excellent*. Since these all contain explicit

mention of the stronger alternative, Pankratz & van Tiel's measure is likely also tapping into the contextual relevance of the alternatives themselves; in fact, Hu *et al.* (2023) also relied on scalar constructions in their operationalization of alternative uncertainty. Further, even though Pankratz & van Tiel's work aims to model "general relevance", it follows the usage-based assumption that this can be approximated by averaging over different individual contexts found in a corpus. Their findings can therefore be interpreted as evidence that the contextual relevance of alternatives matters for (the variation in) SI calculation.

In this paper, we approach the issue of alternative uncertainty by manipulating the Question Under Discussion (QUD, Roberts 1996/2012) prior to a target utterance. An explicit QUD makes the contextually relevant alternatives highly salient to comprehenders and therefore reduces the uncertainty associated with the identity of alternatives. As we will show in our experiments, this indeed substantially reduces the inter-scale variation in SI calculation. It is interesting to note, however, that reducing uncertainty about what a relevant alternative should be does not completely eliminate scalar diversity, raising questions about other sources of uncertainty. We argue that even when relevant stronger alternatives are made salient, the step of excluding those stronger alternatives does not automatically follow. As mentioned, under standard (neo-)Gricean accounts of SI, a hearer reasons about the relevant informationally stronger alternatives and then in the appropriate contexts makes the pragmatic move to exclude those alternatives. But depending on the context, there are other legitimate pragmatic moves that would not require the exclusion of the stronger alternative. That is to say, for the pragmatic calculation of SI, there could be uncertainty associated with both the identity of the relevant alternatives and the necessity of excluding stronger alternatives. To examine whether reinforcing the step of alternative exclusion would also lead to reduced scalar diversity, we will make use of the focus particle *only*, which grammatically requires the exclusion of alternatives (Rooth, 1985, 1992; Krifka, 1999). We show that when both types of uncertainty are removed, scalar diversity can indeed be reduced to the minimum. In the next section, we elaborate on our two experimental manipulations that aim to reduce uncertainty.

2 Pragmatically and semantically based exclusion of relevant alternatives

Context-sensitivity is a hallmark of pragmatic inferences. Indeed, it has long been noted in the theoretical literature that depending on context, SIs are more or less likely to arise (Matsumoto, 1995; Van Kuppevelt, 1996). There is also robust experimental evidence supporting this. For example, Zondervan *et al.* (2008) tested participants' likelihood of calculating the *some but not all* SI from sentences such as *Some pizzas were delivered.* Crucially, these sen-

tences appeared in a dialogue context that either promoted the stronger alternative *all* (3), or the weaker scalar term *some* (4).

(3) A: Were all pizzas delivered?
B: Some pizzas were delivered.

(4) A: Were some pizzas delivered?
B: Some pizzas were delivered.

The authors found a significantly higher rate of SI calculation in the former than in the latter context (43% vs. 7%). Similar results have been obtained using different ways of manipulating context, e.g., by varying an implicit background story (Degen, 2013) or varying intonation (Cummins & Rohde, 2015). Moreover, context has been shown to not only affects hearers' likelihood of calculating an SI, but also the attendant processing cost (Degen, 2013; Degen & Tanenhaus, 2015; Kursat & Degen, 2020; Ronai & Xiang, 2021b) and how adult-like children's SI calculation is (i.a. Papafragou & Tantalou, 2004).

In our study, we capitalize on the context-sensitivity of SI and use explicit QUDs to reduce uncertainty about alternatives, in a way that will be made clear using the following example. Concretely, we make SI-triggering sentences answers to a question that contains the stronger alternative, as in (5).

(5) A: Is the movie excellent?
B: The movie is good.

This context manipulation not only makes a particular stronger alternative (*excellent*) salient in the discourse, but it also encourages SI calculation for reasons having to do with question-answer congruence. Following standard semantic treatments of questions (Hamblin, 1976; Groenendijk & Stokhof, 1984) —which we elaborate on more formally in Section 4.2 —we can take A's question in (5) to set up two possible answers based on the word *excellent*. One is *The movie is excellent* and the other *The movie is not excellent*. What constitutes a good (or in other words, congruent) answer to a question can be defined as one that determines which of the two possible answers the question had set up is actually true (Hulsey *et al.*, 2004; Gualmini *et al.*, 2008). Considering the literal (2-a) and SI-enriched (2-b) readings of the sentence *The movie is good*, only the SI-enriched one (*The movie is good, but not excellent*)

constitutes a good answer, since it entails one of the two possible answers (*The movie is not excellent*). In this way, a context manipulation like (5) encourages SI calculation—that is, hearers' reasoning about and ruling out of alternatives. Indeed, our findings show that this manipulation both makes SI calculation more likely and reduces scalar diversity. At the same time, even in the context of a dialogue like (5), calculating the SI and excluding the stronger alternative is only one potential pragmatic move among many. There are other ways to interpret B's response to A, e.g., as conveying an ignorance inference, paraphrasable as: *I don't know whether the movie is excellent; all I know is that it's (at least) good.* Consequently, as we will show, the QUD manipulation does not completely eliminate inter-scale variation.

As mentioned, hearers therefore face uncertainty not just when it comes to the identity of the relevant stronger alternatives, but also whether those alternatives should be excluded. To reinforce the step of alternative exclusion in the inferencing process, we conduct a manipulation using the focus particle *only*, which encodes alternative exclusion in the semantics, making it an obligatory step. Exclusive focus particles have two meaning components: a positive and a negative. A sentence like (6) conveys that the prejacent proposition is true ((6-a), the positive component) and that alternatives to the prejacent are false ((6-b), negative component).

(6) Mary ate only the cookies.

a. Mary ate the cookies.

b. Mary ate nothing other than the cookies.

That is, such a sentence carries information not just about what Mary ate (cookies), but also the meaning that Mary did not eat anything else from among a set of contextually-determined of alternatives, e.g. {muffin, cinnamon bun, waffle, ...} (i.a. Rooth, 1985, 1992; Krifka, 1999; Coppock & Beaver, 2013).

While here we think of explicit questions and focus as eliminating uncertainty associated with the two steps in inference calculation—identifying alternatives and excluding them—questions and focus are closely related more generally (i.a. Roberts, 1996/2012; Rooth, 1985, 1992; Beaver & Clark, 2008). For instance, it has been observed that an answer to a wh-question has to have a congruent focus structure: the position of the focused item in the answer has to correspond structurally to the position of the wh-word in the question. For this reason, B's answer to A in the following dialogue is infelicitous with focus on the adjective *good*, but felicitous with focus on *the movie* (where focus is marked with capital letters).

(7) A: What was good?
B: #The movie was GOOD.
B′: THE MOVIE was good.

One way this constraint can be formalized in the semantics is by positing that particles such as *only* are grammatically constrained to pick up the QUD (as proposed in Beaver & Clark, 2008, p. 249). This general association between questions and focus also makes them a natural pair of manipulations.

Importantly for our purposes, *only* can associate with scalar terms, for example the weaker scalar *good*:

(8) The movie is only good.

This allows us to demonstrate the critical difference between alternative exclusion encoded by SI vs. *only*, namely the status of that exclusion in the grammar. SI excludes alternatives pragmatically, via a cancellable pragmatic inference: as (9-a) shows, the *not excellent* SI can be cancelled, and the alternative *excellent* ruled back in. In the case of focus with *only*, alternatives are excluded at the level of semantics, and therefore this exclusion cannot be cancelled (9-b). Since the *not excellent* meaning encoded by *only* is not an SI, throughout this paper we adopt the more general term *upper-bounded inference* to refer to the inference arising from sentences like (8).

(9) a. The movie is good. In fact, it's excellent.
b. The movie is only good. #In fact, it's excellent.

Given the above, we expect the focus manipulation to eliminate uncertainty regarding whether alternatives should be excluded. And indeed, in the studies we will present below, we find that *only* (*only good*, *only some*) does increase inference rates and leads to a reduction in scalar diversity to a larger extent than the discourse context manipulation. However, this does not remove all uncertainty: even though the focus particle *only* makes ruling out an alternative obligatory and non-cancellable, it does not specify the identity of the relevant alternative(s). Thus there remains the uncertainty about the relevant alternatives, and consequently, some inter-scale variation. But as we show in our experiments, once both types of uncertainty are removed, i.e. when the identity of alternatives is made clear by an explicit

question and their exclusion is made maximally necessary (being grammatically encoded by focus with *only*), we find ceiling-level uniform calculation of upper-bounded inferences, and the elimination of the scalar diversity effect.

In order to precisely evaluate the potential sources of scalar diversity and how different sources interact, we need a metric that can quantitatively characterize the inter-scale variation. Previous work on scalar diversity has predominantly relied on visual inspection and providing the range of SI rates to assess the variation across lexical scales. We instead propose to use the information theoretic measure of relative entropy to quantify scalar diversity. We are then able to use this measure to assess whether, and how much, the different manipulations (context, *only*, their combination) modulate variation in inference rates. The rest of this paper is structured as follows. In Section 3 we replicate scalar diversity (Experiment 1) and describe our proposal to use relative entropy to quantify it. In Section 4 we conduct the context manipulation (Experiment 2), while in Section 5 we introduce the manipulation with *only* (Experiment 3). In Section 6, we combine these manipulations (Experiment 4). Section 7 offers general discussion and concludes.

3 Experiment 1: Quantifying scalar diversity with relative entropy

Experiment 1 replicates the scalar diversity effect on 60 different lexical scales, and describes our proposal to quantify diversity.

3.1 Data availability

The following OSF repository contains experimental data, stimuli sentences, as well as scripts used for data visualization and statistical analysis, for all experiments reported in this paper: https://osf.io/z3tw2/?view_only=776336a7383248fd9d0ec2faea42f426

3.2 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. Informed consent was obtained from participants. Native speaker status was established via a language background questionnaire, where payment was not conditioned on the participant's response. 1 participant was removed from analysis because the background questionnaire revealed that they were bilingual. 1 additional participant was removed based on having a reaction time shorter than 500ms on the majority of the trials, as well as answering "No" in the first half and "Yes" in the second half of the trials, suggesting that they

were not paying attention to the task. Data from 40 participants is reported below.

Following van Tiel *et al.* (2016) (see also Pankratz & van Tiel 2021), we used an inference task to investigate the likelihood of deriving an SI. Participants were presented with a sentence such as "Mary: *The movie is good.*" and were asked the question "Would you conclude from this that Mary thinks the movie is not excellent?". They responded by clicking "Yes" or "No". Figure 1 shows an example trial item.

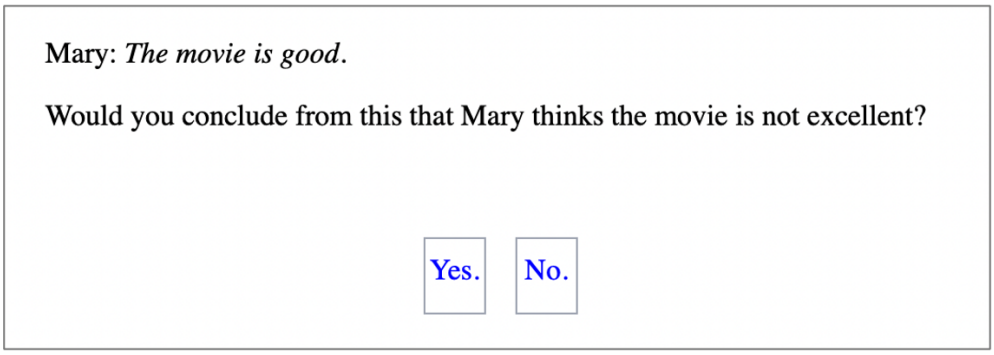

Figure 1: Example experimental trial from Experiment 1.

A "Yes" answer indicates that the participant has calculated the relevant SI (*good* → *not excellent*), while a "No" answer indicates that the participant has not calculated the SI, i.e., they are interpreting *good* as meaning *at least good*, compatible with *excellent*.

60 different lexical scales were tested in the inference task. To identify 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 & 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 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 & Tonhauser (2019). This resulted in a total number of 101 items. As the next step, the following semantic tests were conducted to probe whether X and Y indeed form a scale:

- 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; these judgments were made by the first author and a native speaker consultant. Lastly, wherever one word participated in more than one scale, all but 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.

Overall, the scale collection procedure resulted in a final set of 60 <*weaker, stronger*> scalar terms, which served as our experimental items. Wherever possible (for 33 scales), carrier sentences (e.g., *The movie is good* for <*good, excellent*>) were adopted from van Tiel *et al.* (2016) or de Marneffe & Tonhauser (2019), with minor modifications. For the remaining 27 scales, carrier sentences similar to those used in prior work on scalar diversity were created. The goal in the generation of these carrier sentences was to make them neutral: the noun that the scalar term is predicated of was to be compatible with the literal meaning (e.g., *The sales will at least double*), the SI-enriched meaning (*The sales will double, but not triple*), and the stronger alternative (*The sales will triple*). These judgements were made by the first author and a native speaker consultant[1].

7 filler items from van Tiel *et al.* (2016) were also included, which contained two terms that are either in an entailment relation (*wide* → *not narrow*), or are unrelated to each other (*sleepy* → *not rich*). Given that the filler items had a clear, correct "Yes" or "No" answer, they were included to serve as catch trials. The experiment began with 2 practice trials to familiarize participants with the task; following that, each participant saw 67 trials.

[1]While we believe that this way of creating carrier sentences is in line with existing scalar diversity studies, including a wider variety of (naturally occurring) sentence frames per scale is desirable in future work. Though van Tiel *et al.*'s (2016) scalar diversity experiment used three different sentences per scale and found no differences between them, Aparicio & Ronai (2023) have recently shown that carrier sentences do have a small but significant effect. Further, Degen (2015) and Sun *et al.* (2023) have demonstrated that SI calculation varies across corpus examples —a point we return to in Section 7.3.

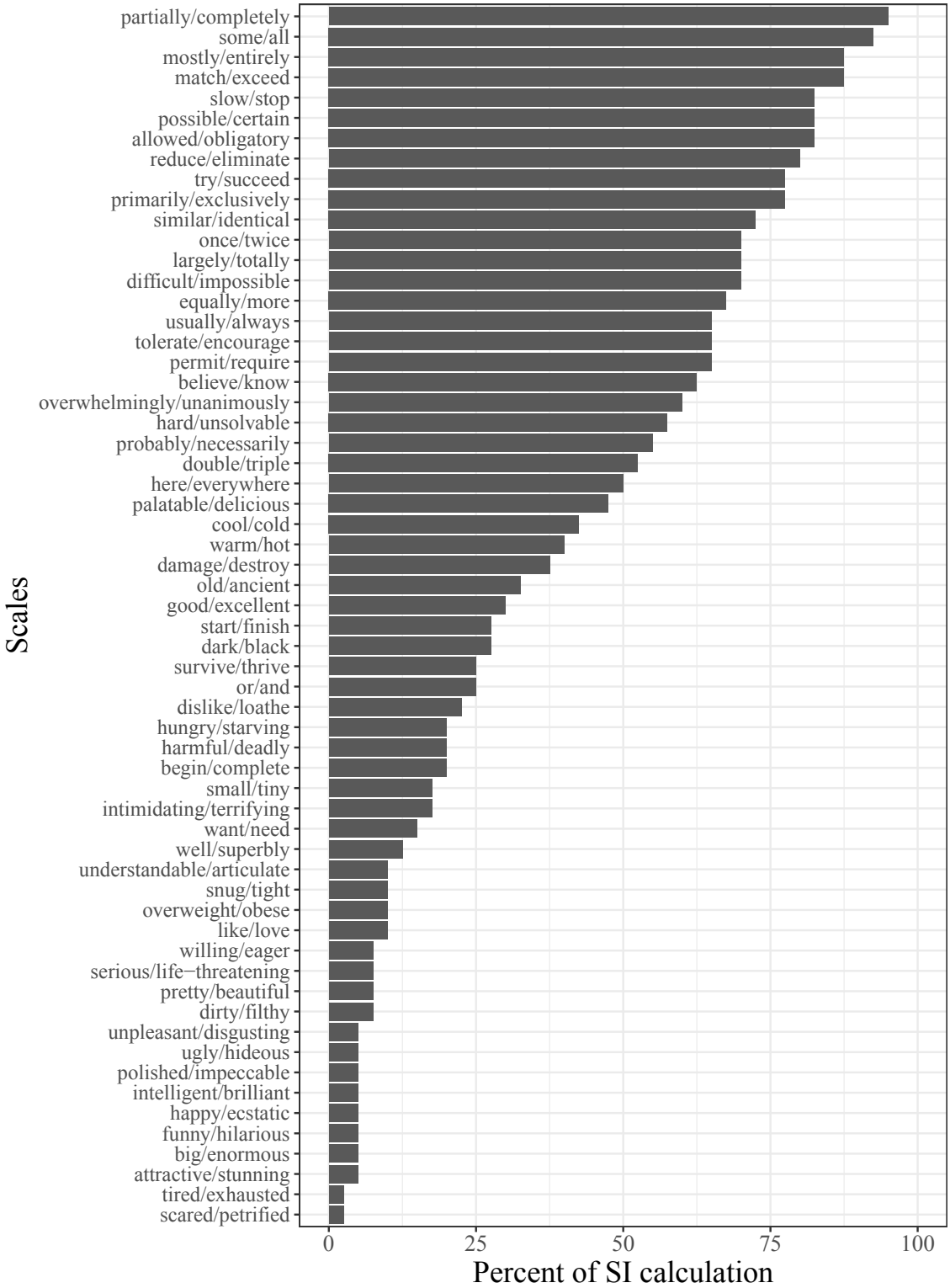

Figure 2: Results of Experiment 1: Inference rates for 60 different scales.

3.3 Predictions

Given consistent findings of scalar diversity in existing literature (reviewed in the Introduction), we predict robust variation across the 60 different scales in how likely they are to lead to SI calculation. That is, we predict that the percentage of "Yes" vs. "No" responses in the inference task of Experiment 1 will vary substantially from scale to scale.

3.4 Results

Figure 2 shows the results of Experiment 1. Percent of inference calculation corresponds to the proportion of "Yes" responses. The average rate of SI calculation across all scales was 38.71%. But as is evident from this figure, considerable variation was found among critical items. In particular, the rate of SI calculation ranged along a continuum from 2.5% (for <*scared, petrified*> and <*tired, exhausted*>) to 95% (for <*partially, completely*>). This result thus successfully replicates the scalar diversity effect: different scalar expressions yielded wildly different rates of SI.

The observation of scalar diversity in prior work was largely based on visual inspection and providing the range of SI rates[2]. Since our goal in this paper is to quantify the effects of various experimental manipulations on scalar diversity, and ranges are a somewhat uninformative statistic, we adopt a more rigorous measure to characterize the variation across scales. For this, 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 quantifying scalar diversity via 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., percentage of "Yes" responses) across different scales as a probability distribution. We then compared this distribution to the uniform distribution. 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 uniform distribution represents a (hypothetical) scenario where each scale leads to the same SI rate. This reflects that the mechanism underlying SI calculation is not typically taken to vary across different scales, leading to the expectation that, all else being equal, different scales lead to SI at the same rate[3]. The uniform distribution operationalizes the uniformity

[2]A notable recent exception is Sun *et al.* (2023), who provide the mean, standard deviation, range and variance, and use Levene's test to compare the variance found in different scalar diversity experiments. We provide a comparable analysis of our own results in the Appendix.

[3]It must be noted that there were some early acknowledgements of non-uniformity in the literature. For

and homogeneity assumptions identified by van Tiel *et al.* (2016) and Degen (2015), respectively. Van Tiel et al. note that, since experimental studies had predominantly tested only a single scale (typically <*some, all*> or <*or, and*>), they must have been operating under the tacit assumption that one scale is "representative for the entire family of scalar expressions" (p. 139) and that "observations about the behavior of a particular lexical scale can typically be generalized" (p. 140). Similarly, focusing on the <*some, all*> scale, Degen reviews (and challenges) the assumption in the semantics/pragmatics literature that implicature strength is invariable, attributing it to earlier work that had viewed SIs as generalized conversational implicatures (Grice, 1967) or default inferences (Levinson, 2000).

The definition given in (10) was used to calculate relative entropy.

(10) 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

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

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 the probability distribution; and $q(x) = 1/60$ is the uniform probability mass function over the 60 scales. Note that we are interested in the relative entropy of a set of experimentally collected inference rates with respect to a hypothetical uniform inference rate. In this specific case, because the uniform inference rate is a constant across all 60 scales, relative entropy can be simplified to the entropy of the uniform distribution minus the entropy of the experimentally collected SI rates[4].

The relative entropy of the SI rates from Experiment 1 is 0.466. To contextualize this number, we may consider a number of hypothetical scenarios as benchmarks. If all scales indeed led to SI calculation at a uniform rate, then that would give a relative entropy of 0—see the Benchmark 1 facet in Figure 3. At the other extreme, the highest possible relative entropy

instance, Horn (1972) makes a distinction between forced and invited inferences (Section 2.15), observing that some SIs "must be inferred by the listener" while others "may be inferred" (p. 112). And, as mentioned at the beginning of Section 2, the context-sensitivity of SI calculation has also been tackled in existing literature.

[4]This means that entropy and relative entropy only differ by a constant (the entropy of the uniform distribution: log(60)) for all our experiments. Nonetheless, we chose to define our measure for diversity as relative entropy, since our goal is to compare experimentally collected data to the uniform distribution, rather than to evaluate the variability found within a single experiment on its own, and relative entropy better reflects this goal. Additionally, this way the relative entropy of each experiment can be interpreted against the benchmark of 0, which is what we would get if all scales in an experiment led to the same SI rate.

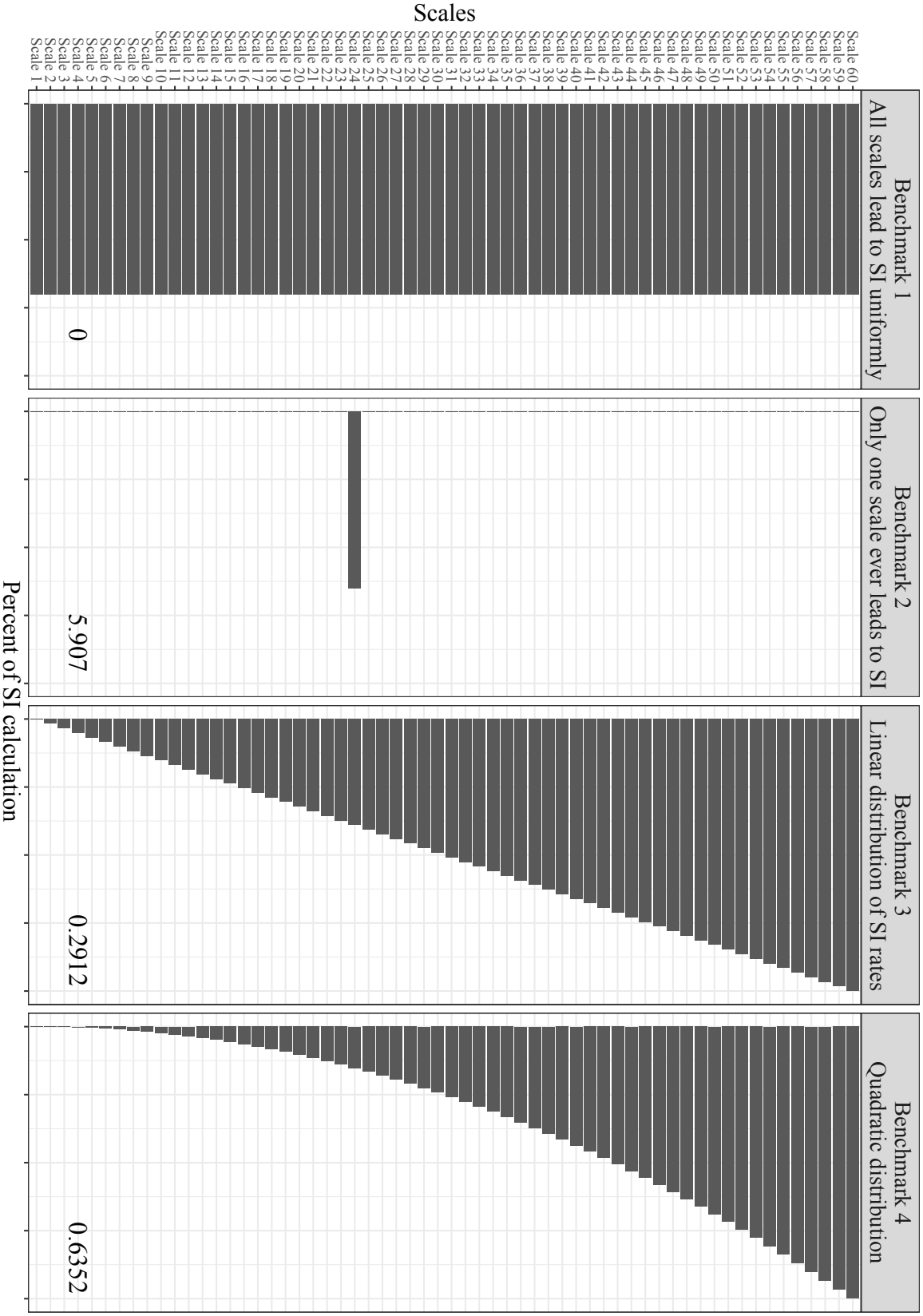

Figure 3: Benchmarks for quantifying diversity using relative entropy: hypothetical SI rates and corresponding relative entropy values (shown in the bottom right corner of each facet).

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 range. 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 end of the distribution, 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.

This demonstrates how the scalar diversity phenomenon can be quantified in a way that goes beyond previous work. We must keep in mind that the benchmarks outlined above are for illustration; the real utility of the proposed relative entropy measure lies in allowing us to compare different sets of experimentally collected SI rates to one another. This is what we turn to in the following sections, where we test the effect of our QUD and focus-*only* manipulations on scalar diversity.

4 Experiment 2: Discourse context manipulation

As outlined in Section 1, in the rest of the paper we explore sources of uncertainty that may lead to scalar diversity, starting with the uncertainty hearers face regarding the identity of contextually relevant stronger alternatives. The question manipulation of Experiment 2 makes these alternatives highly salient, eliminating uncertainty. As mentioned (Section 2), our manipulation follows from prior theoretical and experimental work that has demonstrated the context-sensitivity of SI calculation. But while prior work on the effect of QUDs on SI calculation has largely investigated a single scale such as <*some*, *all*> (with the exception of Cummins & Rohde 2015), here we are interested in the role QUDs might play in making SI calculation more likely across a large variety of different lexical scales, and importantly, in whether they reduce the observed inter-scale variation, i.e., scalar diversity.

4.1 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 either preceded by a polar question that contained the stronger scalar alternative ("strong QUD" condition), or by a polar question that contained the weaker scalar term ("weak QUD" condition). For the <*good*, *excellent*> scale, for instance, the manipulation included the question *Is the movie excellent?*–see Figure 4–, or the question *Is the movie good?*. The question manipulation (strong vs. weak QUD) was administered within participants in a Latin Square design.

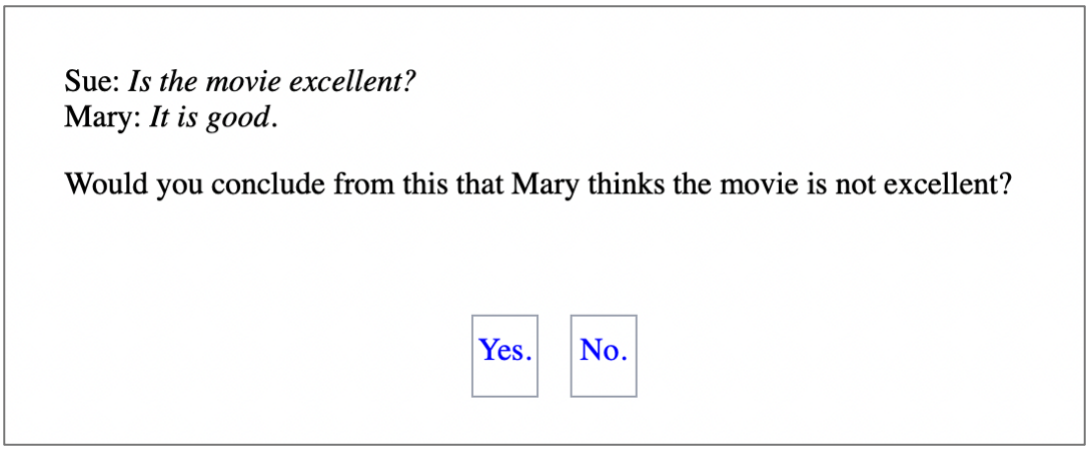

Figure 4: Example experimental trial from Experiment 2

Items (now: Mary's answers) were modified from Experiment 1 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.

4.2 Hypothesis and predictions

Let us revisit in more detail what effect we expect the context manipulation to have, which was already briefly described in Section 2. Sue's question in the strong QUD condition makes the stronger alternative (e.g., *excellent*) highly salient. In other words, the strong QUD condition makes it clear what is a (contextually) relevant alternative to the weaker scalar term (*good*) in Mary's utterance. This is predicted to reduce hearers' uncertainty

about the identity of alternatives and increase robustness of SI calculation. Additionally, if scalar diversity arises, in part, because there is inter-scale variation in uncertainty about alternatives, then the strong QUD manipulation should eliminate this variation—since it makes the alternative clear for all scales— and reduce scalar diversity.

As mentioned, similar predictions follow from the perspective of question semantics and question-answer congruence. Questions partition a set of possible worlds into cells denoting their possible answers (Hamblin, 1976; Groenendijk & Stokhof, 1984). For the strong QUD condition, this results in one cell of the partition containing all the worlds where the movie is excellent, with the other cell containing all the worlds where the movie is not excellent—see (11).

(11) Partitioning from *Is the movie excellent?*

a. Cell 1: The movie is excellent.

b. Cell 2: The movie is not excellent.

In the weak QUD condition, on the other hand, a set of possible worlds is partitioned based on the weaker scalar term *good*: in one cell are all the worlds where the movie is good, and in the other cell, all the worlds where the movie is not good (12).

(12) Partitioning from *Is the movie good?*

a. Cell 1: The movie is good.

b. Cell 2: The movie is not good.

An answer can be taken to be congruent with (or "a good answer to") a question if it determines which of the two resulting cells contains the actual world (see e.g., the Question-Answer Requirement of Hulsey *et al.* 2004; Gualmini *et al.* 2008). Consider now the two readings (literal and SI-enriched) of the potentially SI-triggering sentence *The movie is good* in this light. With the strong QUD, only the SI-enriched meaning (*The movie is good, but not excellent*) is a congruent answer, because it entails the "not excellent" cell of the partition, and eliminates the "excellent" cell. The literal meaning (*The movie is at least good*), on the other hand, does not entail either cell, and it therefore does not directly bear on the question. In the weak QUD condition, however, *The movie is good* constitutes a good answer no matter whether it gets enriched to mean *not excellent*, since it entails the "good" cell of the partition (and eliminates the "not good" cell) in either case. Based on this, we can make the prediction that

strong QUDs should encourage SI calculation: SI calculation is necessary to make Mary's answer a congruent one.

Lastly, it is also possible to think of the strong QUD as encouraging the calculation of the *not excellent* inference as a relevance implicature (Maxim of Relation; Grice 1967). Mary, in response to Sue's question which explicitly mentions *excellent*, chooses not to directly agree or disagree, but rather to offer an alternative (*good*). This might lead participants to infer the negation of *excellent* irrespective of SI, that is, even if <*good, excellent*> did not form a scale. For example, if Sue asked *Is the movie good?* and Mary answered *It is popular*, the inference that the movie is not good arises because Mary avoids directly answering the question, even though *good* and *popular* are not on the same scale.

All of the above lines of reasoning lead to the prediction that the strong QUD condition will produce an increase in SI rates as compared to the baseline Experiment 1, which included no context. And as the likelihood of SI calculation increases across the board for all scales, scalar diversity is also predicted to be reduced, since an across-the-board increase means more scales will cluster close to the ceiling. In the weak QUD condition, on the other hand, there is no reason to predict an increase in SI rates, since that question does not eliminate uncertainty about the identity of contextually relevant alternatives, and Mary's answer is congruent no matter whether it receives an SI interpretation.

4.3 Results

Figure 5 shows the results of Experiment 2 (second facet: "Weak QUD" and third facet: "Strong QUD"), along with the result of Experiment 1 (first facet: "SI"). Averaged over all 60 scales, the rate of SI calculation was 34.53% in the weak QUD and 61.23% in the strong QUD condition. To compare the rates of inference calculation in the weak vs. strong QUD conditions, 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 Condition. It 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 Condition (weak QUD vs. strong QUD) was sum-coded before analysis (weak QUD: -0.5 and strong QUD: 0.5). This analysis revealed that the strong QUD condition led to significantly higher SI rates than the weak QUD condition (Estimate=1.67, SE=0.1, z=16.15, p <0.001).

In the next analysis, we compared the rates of inference calculation in Experiment 2 (in either the weak or the strong QUD condition) to the inference rates from Experiment 1.

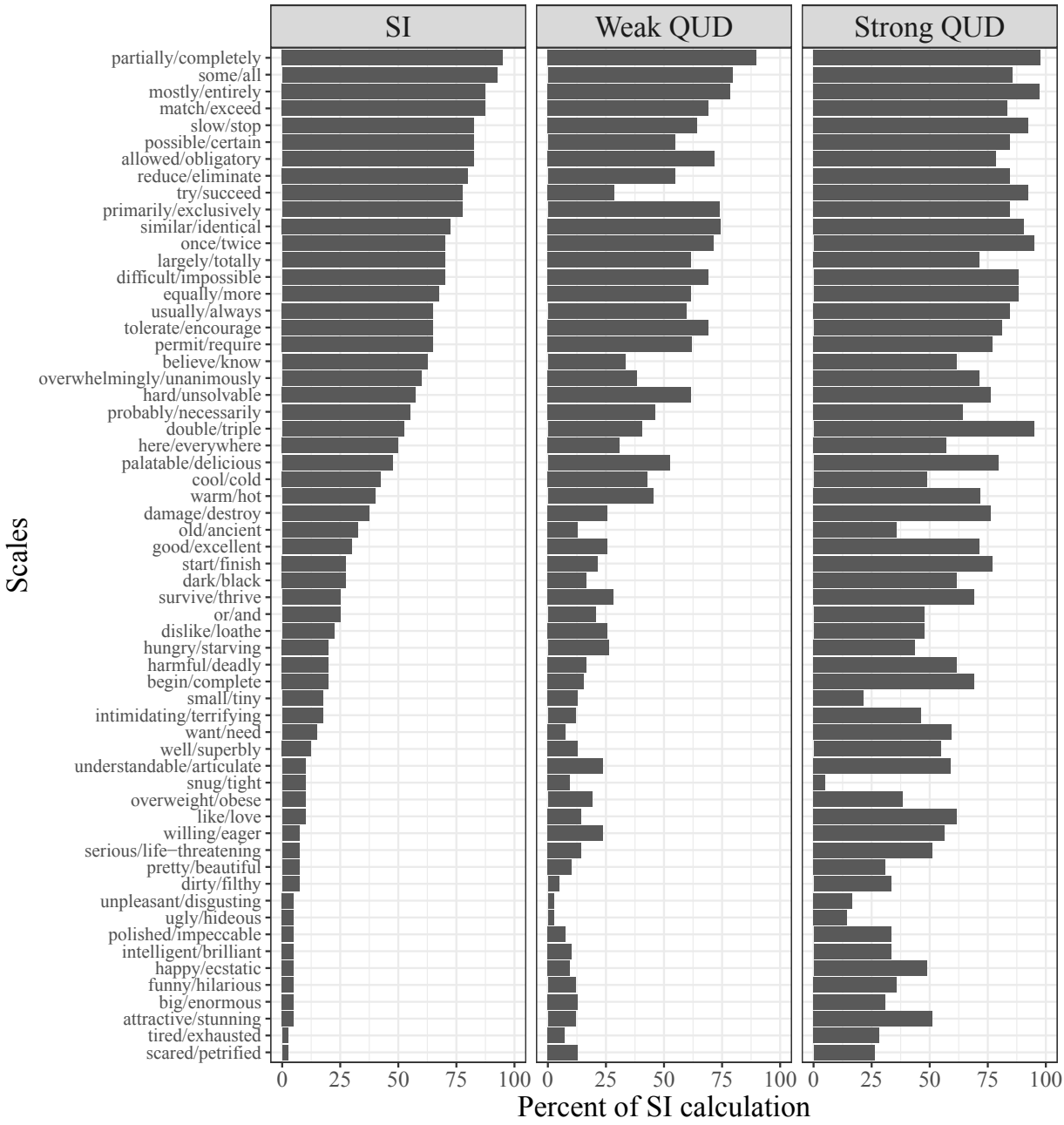

Figure 5: SI rate for 60 different scales. Experiments 1 (SI) and 2 (Weak QUD, Strong QUD) are shown on the three facets of the plot.

To do this, we fit a logistic mixed effects regression model to the combined Experiment 1-2 data set, which predicted Response ("Yes" vs. "No") as a function of Condition (Experiment 1 vs. Experiment 2 weak QUD vs. Experiment 2 strong QUD). The Condition predictor was treatment coded, with Experiment 1 set as the baseline. The random effects structure included random intercepts for participants and random slopes and intercepts for items. This analysis revealed an overall increase in inferences rates in Experiment 2's strong QUD condition, as compared with Experiment 1 (Estimate=1.44, SE=0.21, z=6.88, p <0.001). In the weak QUD condition, however, inference rates were not statistically different from those in Experiment 1 (Estimate=-0.21, SE=0.2, z=-1.1, p =0.28).

To check the effect of the discourse context manipulation on the variation in SI rates across scales, we can turn to relative entropy. The SI rates in the weak QUD condition of Experiment 2 resulted in a relative entropy of 0.378, while the strong QUD condition resulted in a relative entropy of 0.123. 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.

Lastly, in order to check whether the relative order of different lexical scales remains consistent across different manipulations, we calculated rank-order correlations using Kendall's τ_B. We found that SI rates under the weak QUD and strong QUD were significantly correlated by-item (τ_B=0.69, p <0.001), and each of them was also significantly correlated with the no-context SI rates from Experiment 1 (weak QUD-Experiment 1: τ_B=0.77, p <0.001; strong QUD-Experiment 1: τ_B=0.7, p <0.001). This suggests strong similarity among the rankings of different scales across the different experimental conditions.

4.4 Discussion

We found that a supportive discourse context (i.e., strong QUD) made participants significantly more likely to calculate inferences, as compared to either a no-context situation (Experiment 1), or the weak QUD context. Additionally, a supportive context reduced the variation in SI rates: relative entropy is lower in the strong QUD condition of Experiment 2 than in Experiment 1, i.e., there is less scalar diversity. The weak QUD condition did not significantly change the likelihood of SI calculation compared to no context. While it did lead to a slight reduction in inter-scale variation compared to Experiment 1, this effect was much less pronounced than in the strong QUD condition. 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 —but a question based on the weaker scalar

term did not have the same effect.

We take these results as supporting our hypothesis that explicit QUDs can reduce hearers' uncertainty associated with the identity of unsaid alternatives. It must be noted that a key property of the strong QUD manipulation is that it explicitly mentions the relevant stronger alternative, increasing its salience. However, there is reason to believe that simply increasing the alternative's salience, without context, would not have produced the same results. An informative comparison in this regard is between our Experiment 2 and Schwarz *et al.* (2016). Schwarz *et al.* (2016) conducted a priming experiment along these lines, where participants were shown the stronger alternative (as a prime word) before seeing a potentially SI-triggering sentence (as a target). This manipulation, however, did not increase SI rates, nor did it substantially reduce inter-scale variation. That we found an effect of the QUD manipulation, in contrast, suggests that what scalar diversity is modulated by is the contextual relevance of alternatives, and not just their salience *per se*.

At the same time, even with the strong QUD, we did not find a ceiling effect in SI rates, nor did we find uniformity across scales. Instead, SI rates remain relatively low (61.23% across all scales), and they also remain relatively diverse —as we will see in comparison to Experiments 3-4. As previewed in the Introduction, we argue that this is because even when alternative uncertainty is eliminated, there remains uncertainty about whether to exclude that alternative. Instead of alternative exclusion, there are other pragmatic moves hearers (and our participants) may make. Specifically, we propose that there are in fact three different possible pragmatic inferences that can be attributed to Mary's utterance in the dialogue context, which we detail below in (13-a)-(13-c).

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

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

Example (13-a) is the standard SI, which is what arises if hearers both correctly identify the contextually relevant alternative and make the pragmatic step of excluding it. But another possible interpretation that (some) participants in Experiment 2 may have assigned to Mary's answer is the inference in (13-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. For instance, Mary may not know what the threshold is for something to count as excellent (see Kennedy & McNally 2005; Kennedy 2007). (13-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. Perhaps a better illustration of this is a scale like <*snug*, *tight*> (Sue: *Is the shirt tight?*; Mary: *It is snug.*), where the two scalar terms can be more easily construed as each other's (near-)synonyms; correspondingly, for this scale, we found barely any SI calculation even in the strong QUD condition.

Importantly, in addition to the inference in (13-a), (13-c) also represents a congruent answer, as per the definition set out in Section 4.2. While (13-a) addresses the question by entailing "not excellent", (13-c) addresses it by entailing "excellent" —assuming that (13-a) and (13-c) are interpreted on their respective enriched meanings. The availability of the reading in (13-c) —that is, how easily two scale-mates can be construed as synonyms —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. As van Tiel *et al.* (2016) and Ronai & Xiang (2022) have demonstrated, the less distant or distinct the two scalar terms are, the less likely the SI is in a no-context situation (like Experiment 1). Additionally, the less distinct they are, the more it is possible to interpret the weaker term as a synonym for the stronger alternative, as in (13-c), and the lower the SI calculation rate stays even with a biasing strong QUD context (Experiment 2).

Zooming out, as noted in Section 4.2, one crucial observation about the dialogue manipulation in Experiment 2 is that in the strong QUD condition, Mary's answers can be interpreted as indirect answers. That is, the experiment sets up a context where Sue's polar question is answered not with a "Yes" or a "No"[5], but indirectly. It is therefore possible that participants interpreted Mary's answer as implying that the movie is not excellent simply because she did not say "Yes", but not because of an SI-calculation process. However, this same line of reasoning could also apply to the lack of a "No" response from Mary: a hearer might interpret Mary's answer as communicating that the movie is indeed excellent, corresponding to the meaning in (13-c), due to her not responding with "No". It is worth noting too that indirect answers to a polar question are not always perceived as clearly conveying a "Yes" or "No" answer at all; see i.a., de Marneffe *et al.* (2010) and Louis *et al.* (2020) for recent experimental-computational work on how different indirect answers can be interpreted vis-à-vis the Yes/No

[5]Note that interpreting Mary's answer as a negative answer corresponds to responding with "Yes" in the inference task (and an affirmative answer from Mary corresponds to "No" in the inference task), since the task asks about the negated version of Sue's question (*...the movie is not excellent?*).

distinction. Given the findings of this literature and our experimental setup, an interesting follow-up would be to test Experiment 2's items but with a response particle (Yes/No) in Mary's answer and ask participants to rate the naturalness of the resulting dialogue. Results of such an experiment could shed light on whether, in the absence of a response particle, an affirmative vs. negative interpretation is more likely for certain lexical scales, which in Experiment 2 could have given rise to a (13-a) or (13-c) interpretation as a consequence of Mary giving an indirect answer.

To summarize, Experiment 2 demonstrates that eliminating one source of uncertainty (what is the identity of the relevant alternative?) reduces scalar diversity. But another source of uncertainty (is it necessary to exclude that alternative?) still remains, and consequently, so does some inter-scale variation.

5 Experiment 3: Focus manipulation

Under (neo-)Gricean accounts of SI, hearers reason about unsaid stronger alternatives and make the pragmatic move to negate those alternatives. We have argued that there may be uncertainty associated with both of these steps. In Experiment 3, we examine what effect reinforcing the step of alternative exclusion has on the likelihood and diversity of inference calculation. To achieve this, we make use of focus—required by the particle *only*—, which makes the exclusion of alternatives necessary in the semantics, eliminating the uncertainty we saw in Experiment 2 regarding which possible pragmatic move to make.

5.1 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.

Experiment 3 employed the same inference task as the previous two experiments. This time, the additional manipulation conducted was to include the focus particle *only* in the inference-triggering statement. That is, Mary's utterance was e.g., *The movie is only excellent*—see Figure 6 for an example trial. Other than this, the materials and procedure were identical to Experiment 1.

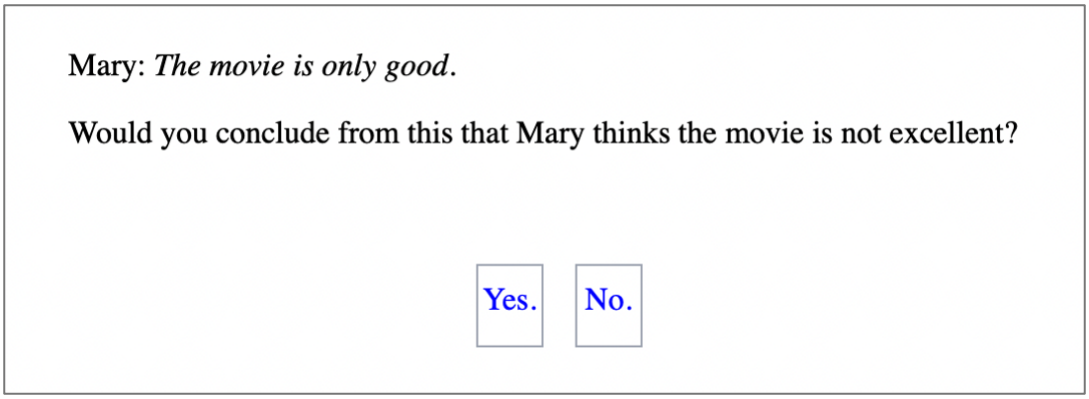

Figure 6: Example experimental trial from Experiment 3

5.2 Hypothesis and predictions

As mentioned in Section 2, *only* shares with SI the property of negating alternatives. In cases where *only* associates with a weaker scalar term (14), the upper-bounded inference this gives rise to is the same as the SI: both (14) and (15) convey that the movie is not excellent.

(14) The movie is only good.
→ The movie is not excellent.

(15) The movie is good.
→ The movie is not excellent.

Importantly, as the below examples (repeated from (9)) demonstrate, alternative exclusion encoded by *only* is semantic, and therefore non-cancellable, in contrast to the pragmatically encoded alternative exclusion in SI, which can be cancelled.

(16) a. The movie is only good. #In fact, it's excellent.
b. The movie is good. In fact, it's excellent.

Based on this, we expect the focus manipulation to eliminate uncertainty—and any potential inter-scale variation therein—regarding whether alternatives should be excluded. Since alternative exclusion is now a necessary step, we predict that comprehenders will robustly exclude alternatives to the focused element, and the rates of upper-bounded inference calculation will consequently increase. As inference rates increase across the board, toward the ceiling, we

also predict that the variation (diversity) observed across lexical scales will be reduced.

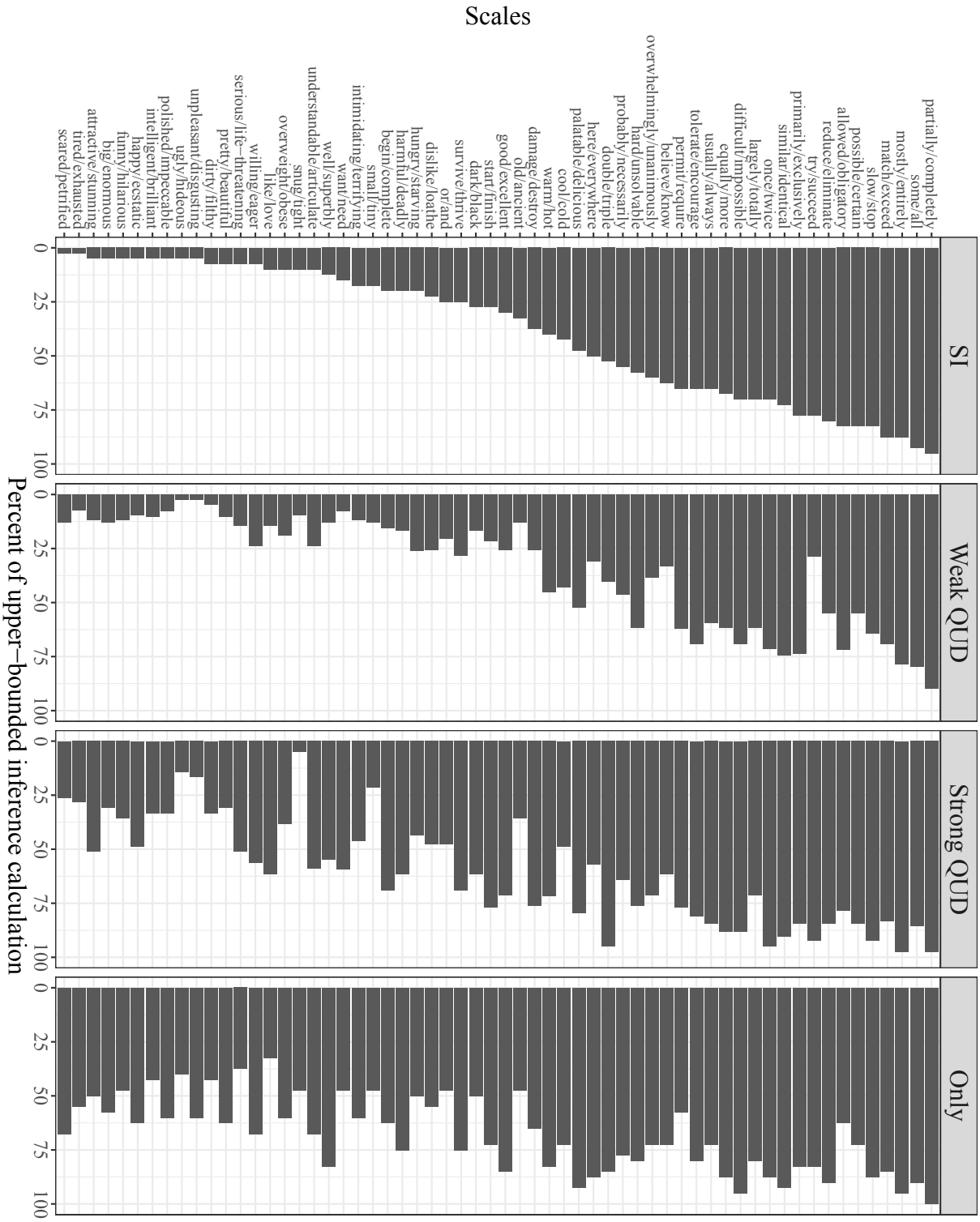

Figure 7: Upper-bounded inference rate for 60 different scales. Experiments 1 (SI), 2 (Weak QUD, Strong QUD) and 3 (Only) are shown on the four facets of the plot.

5.3 Results

Figure 7 shows the results of Experiment 3 (third facet: "Only"), along with the results of Experiments 1 and 2. The average rate of inference calculation across items was 68.42%. 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 4.3. The fixed effects predictor Experiment (1 vs. 3) was sum-coded before analysis (Experiment 1: -0.5 and Experiment 3: 0.5). Random intercepts for participants and random intercepts and slopes for items were included. 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 focus with *only*. However, in a model comparing the Experiment 3 (*only*) results to the strong QUD condition of Experiment 2 (sum-coding: Experiment 2: -0.5 and Experiment 3: 0.5), we find no significant difference (Estimate=0.39, SE=0.26, z=1.52, p =0.13).

Turning 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	Avg. percent of inference
Experiment 1	Baseline scalar diversity	0.466	38.71%
Experiment 2	Weak QUD	0.378	34.53%
Experiment 2	Strong QUD	0.123	61.23%
Experiment 3	Focus with *only*	0.046	68.42%

Table 1: Relative entropy and average inference rate results by experiment: Experiments 1, 2 and 3

Finally, as in the analysis of Experiment 2, we calculated rank-order correlations using Kendall's τ_B. This analysis revealed that inference rates with *only* are significantly correlated with all sets of SI rates from previous experiments: Experiment 1 (SI) - Experiment 3 (*only*) (τ_B=0.54, p <0.001); Experiment 2 (weak QUD) - Experiment 3 (*only*) (τ_B=0.59, p <0.001); Experiment 2 (strong QUD) - Experiment 3 (*only*) (τ_B=0.6, p <0.001). We can observe that these correlations are weaker the more different the average rate of inference calculation gets (e.g., Experiment 1 vs. Experiment 3), and also as the diversity of rates lessens. Nonetheless, all correlations are still significant, suggesting that the relative order of different scales remains consistent under different experimental manipulations.

5.4 Discussion

In line with the predictions, the focus manipulation made upper-bounded inference calculation (*some but not all*, *good but not excellent*) more likely, and it also reduced scalar diversity. Additionally, *only* had a greater effect than the strong QUD in Experiment 2, especially in how much it reduced diversity. This finding makes sense: though both manipulations eliminate uncertainty associated with one of the two steps in inference calculation—identifying alternatives and excluding them—, the question manipulation is fundamentally a pragmatic one. In contrast, grammatically encoded alternative exclusion, i.e., the focus manipulation, constituted a stronger, less violable constraint. However, as is evident from Figure 7, Experiment 3 still did not result in 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. Additionally, there still remains variability across the different scales. We argue that the reason for this is that while the focus manipulation removed the uncertainty associated with the alternative exclusion step, it did not eliminate the uncertainty associated with what can serve as a contextually relevant alternative. We elaborate below.

It has been observed that *only* can draw upon two different "sources" of alternatives; in other words, it is ambiguous between its so-called rank-order reading and its complement-exclusion reading (terminology from Coppock & Beaver 2013, original observation from Horn 1969). The rank-order reading concerns the placement of *good* on a scale where elements are ordered by rank. This reading of *only* can be paraphrased as *no more than*: *The movie is only good* on this interpretation means that the movie is no more than good. On the rank-order reading, then, the excluded alternative must be a stronger one such as *excellent*. A complement-exclusion reading, on the other hand, excludes all alternatives to the focused element, including those that are not ordered on a scale with respect to it. This reading of *only* can be paraphrased as *nothing other than*: *The movie is only good* on this interpretation means that the movie is nothing other than good. Therefore, it is possible to interpret Mary's utterance in Experiment 3 as expressing complement exclusion: communicating that the movie is good, but not other relevant alternatives such as funny or thrilling, etc. Excluding such alternatives leaves open the possibility that the movie is in fact excellent: if the hearer does not take *excellent* to be a relevant alternative to *good* in this particular case, then it will not be ruled out[6]. That is to say, the complement-exclusion reading of *only* does not

[6] A reviewer notes, however, that excluding (all) positive attributes such as *funny* or *thrilling* may still lead to the conclusion that the movie is not excellent, and could in fact be incompatible with the assertion itself (*The movie is good*). This situation would arise if hearers take the criteria for a movie being *excellent* (or even *good*) to be the complement-exclusion alternatives that are being excluded. As the reviewer suggests, however, in the right context, complement-exclusion readings of *The movie is only good* can still be compatible with *excellent*, as the below example demonstrates:

necessarily lead to an upper-bounded inference; only the rank-order reading does[7].

Support for the idea that there remains uncertainty regarding the identity (and in fact type) of alternative under *only* comes from a comparison with other focus particles. While *only* readily admits both rank-order and complement-exclusion readings, other particles prefer (or require) one or the other. For instance, Coppock & Beaver (2013) argue that *merely* prefers the rank-order reading. If what underlies the remaining scalar diversity in Experiment 3 is that participants were able to interpret *only* as excluding complement-exclusion alternatives other than *excellent*, then we should see higher rates of upper-bounded inference calculation and less inter-scale variation with *merely*. Ronai & Fagen (2022) report on experiments that provide evidence for this: given *The movie is merely good*, *excellent* is more unambiguously taken to be the alternative that is excluded.

The persistent non-calculation of upper-bounded inferences with *only*, which occurred more with certain scales than others, is informative with respect to the scalar diversity observation within the SI domain. As Figure 7 shows, with the *only* manipulation, some scales are closer to ceiling level inference calculation than others. We have argued that wherever inference rates remained low with a scale, participants likely excluded an alternative other than the rank-order scalar alternative (e.g., *excellent*) the task question probed. In other words, it seems that for these scales, hearers do not easily converge on the particular lexical scale tested (<*good*, *excellent*>). Given the significant correlation between Experiment 1 and Experiment 3, it is plausible that the same factor could underlie (in part) hearers' non-calculation of SIs. In this way, the focus particle manipulation indirectly identifies how available (van Tiel *et al.*, 2016) or accessible (Ronai & Xiang, 2022) the stronger scalar alternative is given a weaker term, and the inter-scale variation therein, which these prior studies had tested more directly using cloze tasks. Relatedly, *only* may also tap into the polysemy of scalar terms, which Sun *et al.* (2018) have identified as a potential predictor of scalar diversity. As the authors note, some weaker scalar terms belong to more than one scale —e.g., a stronger alternative to *hard*

(i) The original movie was good in some ways and bad in others. The sequel is only good. In fact, it's excellent.

Additionally, other scales pattern differently: for instance, on its complement-exclusion reading, *The shirt is snug* may exclude alternatives such as *red*, *patterned* or *knitted*, the exclusion of which would not lead to the conclusion that the shirt is not *tight* either. We tentatively suggest that it is more easily possible for the exclusion of complement-exclusion alternatives to amount to the exclusion of a rank-order one in the case of subjective lexical scales than non-subjective ones, but leave further exploration of this question to future research.

[7]It must be noted that there are uniform formal semantic treatments of rank-order and complement-exclusion readings, e.g., Coppock & Beaver (2013); Beaver & Clark (2008). Nonetheless, what matters for our purposes is that there are still two distinct readings of *only* that correspond to the ruling out of (potentially) distinct alternatives.

is *unsolvable* when it comes to problem-solving but *unbearable* when it comes to suffering (p. 3). In these cases, if the sentential context allows, *only* can be interpreted as excluding a rank-order alternative that is different from the one our experiment tested.

Lastly, related to the above point and to the ambiguity of *only* more generally, we have argued that contextual relevance constrains the space of possible alternatives. But in Experiment 3, stimulus sentences appeared without any context, leaving open the possibility that participants had different contexts in mind. Compare, for instance, (17) and (18).

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

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

As discussed in relation to Experiment 2, a context like (17) removes uncertainty regarding the identity of the alternative; in this case, *only* is most naturally interpreted as excluding this alternative *excellent*. If a participant supposed such a context, they would arrive at the upper-bounded *good but not excellent* inference. But if they supposed a context like the one in (18), 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.

Altogether, while the focus particle *only* encodes alternative exclusion in the semantics, thereby eliminating uncertainty regarding this step in the inferencing process, it does not reduce the uncertainty associated with identifying relevant alternatives. Consequently, in Experiment 3, while scalar diversity was greatly reduced, there still remained some variation[8]. In our final experiment, we therefore turn to a manipulation that eliminates both potential sources of uncertainty—what the contextually relevant alternatives are and whether the step should be taken to exclude them—, by testing dialogues such as (17).

[8]For a small number of 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 inference: *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.

6 Experiment 4: Manipulating both context and focus

In Experiments 2-3, we have seen that eliminating uncertainty either regarding the identity of alternatives (via an explicit QUD manipulation) or regarding the necessity of excluding those alternatives (via a focus manipulation) substantially reduces inter-scale variation in inference calculation. However, in neither case was this variation completely eliminated, which we have argued is due to one kind of uncertainty still remaining: under a QUD manipulation, there are still possible pragmatic moves available other than alternative exclusion, while under a focus manipulation, there is still ambiguity in what (kind of) alternatives should be excluded. In Experiment 4, we combine the previous two manipulations and examine the robustness of inference calculation and the extent of inter-scale variation in a situation where uncertainty is reduced regarding both steps in the calculation process.

6.1 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 manipulations 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 ("strong QUD" from Experiment 2). Figure 8 shows an example trial. Otherwise, the task and materials were identical to previous experiments.

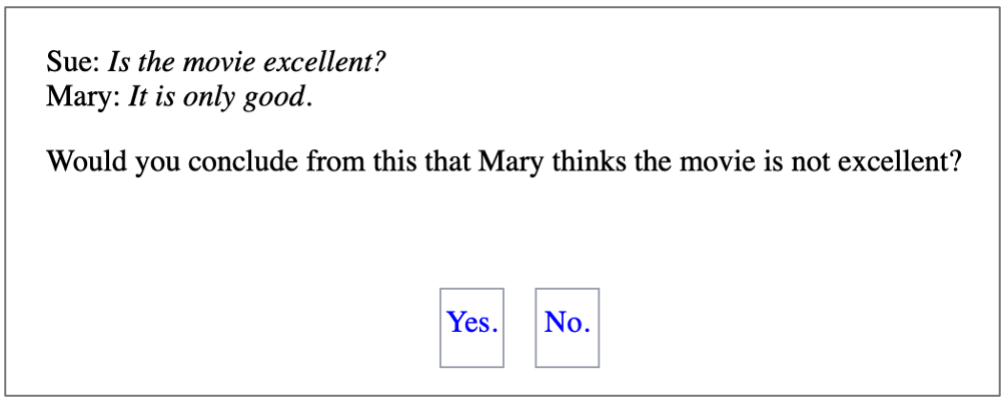

Figure 8: Example experimental trial from Experiment 4

6.2 Hypothesis and predictions

The predictions made for Experiments 2 (Section 4.2) and 3 (Section 5.2) straightforwardly carry over to Experiment 4. First, Sue's explicit questions make the relevant alternatives highly salient, thereby eliminating hearers' uncertainty about their identity, which could have arisen based on Mary's utterance alone. The question manipulation also encourages SI calculation due to question-answer congruence considerations. Second, the presence of *only* in Mary's answer reinforces the step of alternative exclusion. With the identity of alternatives made clear and their exclusion made necessary by the semantics, we predict robust, ceiling-level rates of upper-bounded inference calculation. Since any inter-scale variation in uncertainty should also be eliminated, we also predict that scalar diversity will be reduced to a minimum.

6.3 Results

Figure 9 shows the results of Experiment 4 (rightmost facet: "QUD + only"), along with all previous experiments. As can be seen in the figure, inference rates are now in fact almost at ceiling; the average rate of upper-bounded inference calculation across scales was 88.63%. A statistical analysis identical to the one reported in Section 4.3 was conducted. The fixed effects predictor Experiment (1 vs. 4) was sum-coded before analysis (Experiment 1: -0.5 and Experiment 4: 0.5). Random intercepts for participants and random intercepts and slopes for items were included. This analysis 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). In an additional model, we also compared Experiment 4 (*only* + QUD) to Experiment 3 (*only*). The Experiment predictor was again sum-coded (Experiment 3: -0.5 and Experiment 4: 0.5). The model found significantly higher rates of inference calculation in Experiment 4 (Estimate=1.95, SE= 0.36, z=5.39, p <0.001).

The relative entropy resulting from Experiment 4 is 0.006—see Table 2 for a comparison of the relative entropy from all experiments. We can see that there is little appreciable variation in inference rates across the lexical scales tested.

Lastly, we calculated the Kendall's τ_B rank-order correlations comparing Experiment 4's findings to all previous experiments. This analysis revealed significant correlations between Experiment 4 and each of the previous sets of inference rates: Experiment 1 (SI) - Experiment 4 (QUD + *only*) (τ_B=0.36, p <0.001); Experiment 2 (weak QUD) - Experiment 4 (QUD + *only*) (τ_B=0.4, p <0.001); Experiment 2 (strong QUD) - Experiment 4 (QUD + *only*) (τ_B=0.48, p <0.001); Experiment 3 (*only*) - Experiment 4 (QUD + *only*) (τ_B=0.5, p <0.001).

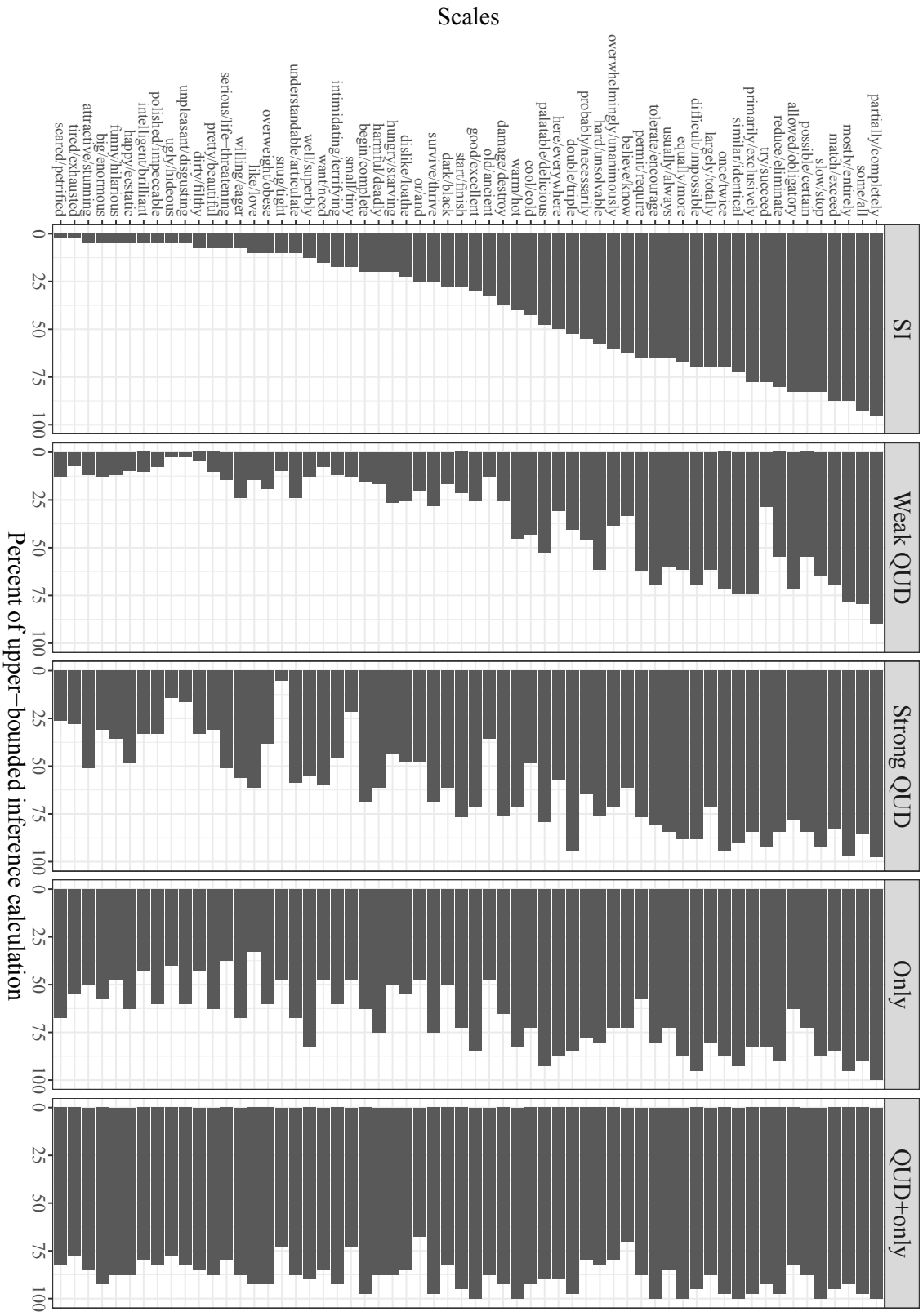

Figure 9: Upper-bounded inference rate for 60 different scales. Experiments 1 (SI), 2 (Weak QUD, Strong QUD), 3 (Only) and 4 (QUD + only) are shown on the five facets of the plot.

	Manipulation	Relative entropy	Avg. percent of inference
Experiment 1	Baseline scalar diversity	0.466	38.71%
Experiment 2	Weak QUD	0.378	34.53%
Experiment 2	Strong QUD	0.123	61.23%
Experiment 3	Focus with *only*	0.046	68.42%
Experiment 4	Strong QUD + *only*	0.006	88.63%

Table 2: Relative entropy and average inference rate results by experiment: Experiments 1, 2, 3 and 4

We note that though we still found highly significant correlations between the relative order of scales across experiments, these correlations are less strong than before, i.e., the coefficients are lower. This is expected, since in Experiment 4 inference rates are near uniform and close to ceiling, which is a stark difference from e.g., Experiment 1. In a situation where there is not much variability across scales, we also would not expect their relative order to match the order found when there was substantially more inter-scale variability.

The implications of these results will be discussed in the next section, along with the rest of our experiments.

7 General discussion

We first review our experimental findings in light of the two sources of uncertainty that we identified (Section 7.1). We then turn to other aspects of the linguistic signal that could potentially eliminate further remaining uncertainty (Section 7.2). Lastly, we discuss how our work allows us to connect SI with relevance implicature and semantic exclusion, as well as how it relates to previous scalar diversity studies and findings of intra-scale variation (Section 7.3).

7.1 Two sources of uncertainty in SI calculation

Though experimental work on SI has tended to focus on a limited number of lexical scales, most commonly <*some*, *all*>, attention has more recently shifted also to the phenomenon of scalar diversity. A number of studies have now shown that, contra the uniformity assumption, different lexical scales (e.g., <*some*, *all*> vs. <*good*, *excellent*> vs. <*intelligent*, *brilliant*>) vary greatly in their likelihood of leading to SI calculation. Our paper was primarily concerned with the as-yet unanswered question of what can explain the robust inter-scale variation. Specifically, we followed up on the hypothesis that (variation in) hearers' uncertainty about what is a relevant alternative can explain the variation in SI rates (i.a. Hu *et al.*,

2023; van Tiel *et al.*, 2016; McNally, 2017). We have further explored the open question of what may be the source of such uncertainty.

In particular, we have proposed that uncertainty may be associated with either of the two steps in a (neo-)Gricean view of the SI calculation process: 1) identifying and reasoning about stronger alternatives that the speaker could have said, but did not, and 2) negating those alternatives. First, we looked at the possibility that contextual relevance serves as a constraint on possible alternatives and therefore modulates scalar diversity. Second, we have argued that there is a second potential kind of uncertainty impacting SI calculation: uncertainty regarding whether the identified alternatives should be excluded. We have approached the issue of these two types of uncertainty by manipulating discourse context via an explicit QUD, making contextually relevant alternatives salient, and by manipulating semantic focus, reinforcing the alternative exclusion step. Our findings revealed that both a supportive explicit QUD and alternative exclusion via *only* led to increased rates of inference calculation. They also both reduced variation (i.e., scalar diversity) —with *only*'s effects being stronger than the discourse context's. To quantify changes in inter-scale variation, we relied on the information-theoretic measure of relative entropy.

Further, we saw that under either manipulation, scalar diversity was not fully eliminated, for which we offered the following explanations. Discourse context can provide salient alternatives, but that alone does not tell hearers that they need to reason about and exclude those alternatives. Instead, it is possible to make different pragmatic moves and assign different interpretations, concluding not that an alternative is false, but e.g., that the speaker does not know whether it is true. The focus particle *only*, on the other hand, makes alternative exclusion obligatory. However, it leaves uncertainty about the identity of the alternative that was to be excluded, allowing hearers to potentially rule out alternatives other than the stronger scale-mate. As our final experiment demonstrated, once both types of uncertainty are removed – the identity of the alternatives is made clear, and the cue to exclude them is encoded semantically – we find ceiling-level inference rates and minimal scalar diversity. When either variety of uncertainty is still present, there is more flexibility in interpretation, and consequently we observe more variation. This also leaves more opportunity for other factors (e.g., distinctness, extremeness, etc., as reviewed in Section 1) to influence the likelihood of inference calculation.

7.2 Further ways of reducing uncertainty

There other aspects of the linguistic signal that we did not manipulate that could also play a role in eliminating uncertainty, going beyond the QUD and *only*. In Section 5.4, we discussed that while *only* leaves uncertainty about the type of alternative to exclude (complement-exclusion or rank-order), other focus particles do not. For example, *merely* better constrains the space of alternatives to rank-order or scalar ones, thereby reducing uncertainty both about the identity of alternatives and whether to exclude them. Indeed, as Ronai & Fagen (2022) have shown, hearers are even more likely to calculate upper-bounded inferences such as *not excellent* from a sentence like *The movie is merely good* than from *The movie is only good*. This is because with *merely*, (non-scalar) complement-exclusion alternatives (e.g., *funny*, *thrilling*) are not as easily understood as the targets of exclusion.

And while we have argued that the QUD manipulation leaves uncertainty about whether identified alternatives should be excluded, some of this could plausibly be reduced by, for instance, intonational cues. Specifically, while (19) (repeated from (13)) allows for three different inferences, only one of which is SI, this would less obviously be the case in a spoken dialogue: Mary's *It is good* answer would likely correspond to different prosodic contours in each of the three cases.

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

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

In particular, the ignorance inference (19-b) might correspond to the so-called rise-fall-rise contour, which Ward & Hirschberg (1985) analyze as conveying uncertainty relative to a scale, and Constant (2012) as alternatives being unclaimable. For *good* to be interpreted as a synonym to *excellent* (19-c), on the other hand, *good* would need to be intensified—for suggestions on what the acoustic correlates of intensification might be, see i.a., Kohler (2006); Niebuhr (2010). Therefore, while Sue's explicit question only reduces the uncertainty associated with the identity of the relevant alternatives, if spoken, Mary's answer would likely additionally reduce uncertainty about which possible pragmatic move (from among (19-a)-(19-c)) is meant.

In fact, inroads have already been made into experimentally investigating the interaction of

prosody and scalar diversity. De Marneffe & Tonhauser (2019) found that the rise-fall-rise contour, contra predictions made by the above-mentioned theoretical accounts, in fact leads to an increase in SI rates —though this paper only reported aggregated results, leaving open the question of whether there is inter-scale variability in the effect of prosody. Evidence for such variability comes from Cummins & Rohde (2015). These authors did not test complex intonational contours such as rise-fall-rise, but found that focus placement on the scalar term (e.g., *good* in *The movie is good*) increases SI rates compared to neutral intonation; however, this effect varied in its strength and was not present for all scales tested. Altogether, such findings make the prosody-scalar diversity interface a promising avenue of further research.

7.3 Broader implications and future directions

The two steps of SI calculation we have tested —the identification and exclusion of alternatives —are relevant not just to SI, but the other phenomena our paper has touched upon, namely relevance implicature and semantic focus. With Experiment 2's QUD manipulation, it is possible that participants derived the target inference as a relevance implicature, while in Experiments 3-4, that inference was encoded by focus with *only*. There are of course important differences across the three types of inference: relevance implicature differs from SI in that the alternatives are contextually provided and do not come from lexical scales, while in the case of semantic focus, alternative exclusion is not a cancellable pragmatic inference. Nonetheless, alternatives and their exclusion are at the core of all these inferences, and our experimental findings have shown that they are also alike in giving rise to the same by-item variation. As demonstrated by significant correlations, the relative order of different scales is consistent across different manipulations (baseline SI, QUD, *only*). That is, when a weaker scalar term was more or less likely to convey the exclusion of a stronger alternative, this remained the case (relatively speaking) across all our experiments. Such a finding leads us to the conclusion that "scalar diversity" —the variation across lexical scales in their propensity to give rise to upper-bounded inferences —plays a role even outside the domain of SI. This, in turn, highlights a further similarity across the three types of alternative-sensitive inferences.

The finding that the relative order of lexical scales remains the same across experiments is also consistent with previous investigations of scalar diversity. Work following van Tiel *et al.* (2016) has tended to use their inference task, or variations thereof, and showed that the relative order of scales consistently replicates —even with e.g., a gradient, rather than binary measure, as in Sun *et al.* (2018, p. 6). Simons & Warren (2018) used a different experimental paradigm to test inter-scale variation, one which placed scalar terms in richer contexts and probed SI calculation without directly presenting participants with the stronger alternatives.

Yet the authors note that they observed very similar patterns to van Tiel *et al.* (2016), which they take to suggest that "a scalar's relative rate of strengthening is quite robust" (p. 277). These prior observations are strengthened by our finding that the relative ordering of scales remains consistent even when context makes alternatives salient or when alternative exclusion is semantically encoded. Whatever factors (e.g., the distinctness of the scale-mates) underlie scalar diversity seem to contribute to the persistent inter-scale variation even in the presence of an explicit QUD or *only*.

Our way of probing what role the two identified sources of uncertainty play in SI calculation might be potentially perceived as different from the predominant method of existing scalar diversity studies. We manipulated the discourse context to eliminate uncertainty about the identity of alternatives and made use of *only* to reinforce the alternative exclusion step, then checked what effect these manipulations had on the variation in inference rates. Most other scalar diversity studies first identified some (typically lexico-semantic) property of the different scales, measured it, and then tested whether the variation in that property across scales can predict the observed variation in SI rates across scales. While these experimental manipulations may seem different on the surface, we would like to highlight that they could be mutually informative and eventually tap into the same underlying processes. For instance, one relevant variable probed by many previous studies is the availability of the stronger alternative given a weaker term on a particular scale (van Tiel *et al.*, 2016; Ronai & Xiang, 2022; Hu *et al.*, 2022, 2023). In the current study, while we did not directly measure this property, our findings based on the focus particle *only* (Experiment 3) converge with previous studies. In particular, the *only* manipulation targets the uncertainty about the exclusion step, but not the uncertainty about what the relevant alternatives are. As noted in Section 5.4 (and 7.1), there are scales that remain unlikely to trigger inference calculation even with *only*, likely due to the remaining uncertainty about what alternatives to exclude. This result, therefore, speaks to the effect of the availability of the strong scalar alternatives. It also suggests that the focus operator could potentially be used as an effective tool to identify those specific scales for which the particular stronger alternatives are less available (van Tiel *et al.*, 2016), accessible (Ronai & Xiang, 2022), or expected (Hu *et al.*, 2022, 2023), due to e.g., the weaker term's polysemy (Sun *et al.*, 2018).

The experiments reported in this paper tested the phenomenon of scalar diversity using stimuli that were manually constructed by the authors. While this is in line with much of the existing experimental pragmatics literature, it is worth addressing how our findings might generalize to more naturalistic language use. An interesting study in this regard is Degen (2015), who tested the *some but not all* SI in a corpus of 1363 naturally occurring sentences

and found substantial variation in the robustness of SI calculation. The approach taken by Degen was recently applied to different lexical scales by Sun *et al.* (2023), who tested SI calculation in a corpus of Twitter data and found reduced scalar diversity compared to van Tiel *et al.* (2016) and Sun *et al.* (2018). Taking these two corpus-based studies together, the picture that emerges is that in naturalistic data, even a single scale such as <*some*, *all*> shows non-uniform SI calculation, and at the same time, inter-scale variation is attenuated (though still clearly present). This raises the possibility that intra- and inter-scale variation might be underlyingly the same. Sun *et al.* (2023) also found that the same factors that predicted scalar diversity in prior experiments (van Tiel *et al.*, 2016; Sun *et al.*, 2018) continue to be significant predictors in their corpus data, and they in fact explain a similar amount of the observed variance. Based on this, one might expect to find an effect of the two kinds of uncertainty tested in our experiments even in more naturalistic settings. Further, as reviewed in the Introduction, Hu *et al.* (2023) have shown that uncertainty about alternatives plays a role in both intra- and inter-scale variation, which would suggest that our own findings might extend to intra-scale variation too. All these findings suggest two big-picture questions that might guide future work on scalar diversity. First, to what extent are scalar diversity and variation within a single scale, at their core, the same phenomenon? And second, how much of the variation in pragmatic inferences arises from the identity of lexical scales vs. contextual cues that are present in natural language data? (See also Degen 2021, who raises similar issues.)

8 Conclusion

In sum, this paper investigated scalar diversity: the robust inter-scale variation in SI calculation. Our experiments add to the understanding of what underlies this variation by investigating two sources of uncertainty as potential explanations. We argued that in the calculation of SI, hearers might have uncertainty regarding the identity of the alternative that is to be excluded, as well as about the exclusion step itself. Our findings showed that scalar diversity is reduced —as quantified by relative entropy —once these sources of uncertainty are eliminated.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. #BCS-2041312. For helpful discussion and feedback, we would like to thank the editor and anonymous reviewers, Itamar Francez, Lucas Fagen, Chris Kennedy, and Michael Tabatowski. We are grateful to Zsolt Veraszto for his help with the discussion on relative entropy,

and to Thomas Sostarics for his help with the discussion on Levene's test and Kendall's τ_B. All remaining errors are due to the authors.

References

Aparicio, Helena, & Ronai, Eszter. 2023. Scalar implicature rates vary within and across adjectival scales. *Pages 110–130 of:* Kim, Juhyae, Öney, Burak, Zhang, Yao, & Zhao, Fengyue (Lisa) (eds), *Proceedings of Semantics and Linguistic Theory (SALT) 33*.

Baker, Rachel, Doran, Ryan, McNabb, Yaron, Larson, Meredith, & Ward, Gregory. 2009. On the Non-Unified Nature of Scalar Implicature: An Empirical Investigation. *International Review of Pragmatics*, **1**(2), 211–248.

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

Bates, Douglas, Mächler, Martin, Bolker, Ben, & Walker, Steve. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, **67**(1), 1–48.

Beaver, David I., & Clark, Brady Z. 2008. *Sense and Sensitivity.* Wiley-Blackwell.

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

Constant, Noah. 2012. English rise-fall-rise: a study in the semantics and pragmatics of intonation. *Linguistics and Philosophy*, **35**, 407–442.

Coppock, Elizabeth, & Beaver, David I. 2013. Principles of the Exclusive Muddle. *Journal of Semantics*, **31**(3), 371–432.

Cummins, Chris, & Rohde, Hannah. 2015. Evoking Context with Contrastive Stress: Effects on Pragmatic Enrichment. *Frontiers in Psychology*, **6**, 1779.

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

de Marneffe, Marie-Catherine, & Tonhauser, Judith. 2019. Inferring Meaning from Indirect Answers to Polar Questions: The Contribution of the Rise-Fall-Rise Contour. *Pages 132–163 of:* Zimmermann, Malte, von Heusinger, Klaus, & Onea, Edgar (eds), *Current Research in the Semantics/Pragmatics Interface*, vol. 36, Questions in Discourse. Leiden, The Netherlands: Brill.

de Marneffe, Marie-Catherine, Manning, Christopher D., & Potts, Christopher. 2010. "Was It Good? It Was Provocative." Learning the Meaning of Scalar Adjectives. *Pages 167–176 of: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics.

Degen, Judith. 2013. *Alternatives in Pragmatic Reasoning*. Ph.D. thesis, University of Rochester.

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

Degen, Judith. 2021. *Harnessing the linguistic signal in predicting within-scale variability in scalar inferences*. Talk presented at the "Scales, degrees and implicature: Novel synergies between semantics and pragmatics" Workshop, https://www.uni-potsdam.de/fileadmin/projects/gotzner-spa/Kickoff_Workshop/Slides_Degen.pdf.

Degen, Judith, & Tanenhaus, Michael K. 2015. Processing Scalar Implicature: A Constraint-Based Approach. *Cognitive Science*, **39**(4), 667–710.

Degen, Judith, & Tanenhaus, Michael K. 2016. Availability of Alternatives and the Processing of Scalar Implicatures: A Visual World Eye-Tracking Study. *Cognitive Science*, **40**(1), 172–201.

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

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

Gotzner, Nicole, Solt, Stephanie, & Benz, Anton. 2018. Scalar Diversity, Negative Strengthening, and Adjectival Semantics. *Frontiers in Psychology*, **9**, 1659.

Grice, Herbert Paul. 1967. Logic and Conversation. *Pages 41–58 of:* Grice, Paul (ed), *Studies in the Way of Words*. Harvard University Press.

Groenendijk, Jeroen, & Stokhof, Martin. 1984. On the Semantics of Questions and the Pragmatics of Answers. *Pages 143–170 of:* Landman, Fred, & Veltman, Frank (eds), *Varieties of Formal Semantics: Proceedings of the Fourth Amsterdam Colloquium*. Foris.

Gualmini, Andrea, Hulsey, Sarah, Hacquard, Valentine, & Fox, Danny. 2008. The Question-Answer Requirement for scope assignment. *Natural Language Semantics*, **16**(3), 205–237.

Hale, John. 2003. The Information Conveyed by Words in Sentences. *Journal of Psycholinguistic Research*, **32**(2), 101–123.

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

Horn, Laurence R. 1969. A presuppositional analysis of *only* and *even*. *Pages 98–107 of: The Fifth Regional Meeting of the Chicago Linguistics Society (CLS 5)*.

Horn, Laurence R. 1972. *On the semantic properties of logical operators in English*. Ph.D. thesis, UCLA.

Hu, Jennifer, Levy, Roger, & Schuster, Sebastian. 2022. Predicting scalar diversity with context-driven uncertainty over alternatives. *Pages 68–74 of: Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*.

Hu, Jennifer, Levy, Roger, Degen, Judith, & Schuster, Sebastian. 2023. Expectations over unspoken alternatives predict pragmatic inferences. *Transactions of the Association for Computational Linguistics*. To appear.

Hulsey, Sarah, Hacquard, Valentine, Fox, Danny, & Gualmini, Andrea. 2004. The Question-Answer Requirement and scope assignment. *Pages 71–90 of:* Csirmaz, Aniko, Gualmini, Andrea, & Nevins, Andrew (eds), *MIT Working Papers in Linguistics*. MITWPL.

Kennedy, Christopher. 2007. Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and philosophy*, **30**, 1–45.

Kennedy, Christopher, & McNally, Louise. 2005. Scale Structure, Degree Modification, and the Semantics of Gradable Predicates. *Language*, **81**(2), 345–381.

Kohler, Klaus J. 2006. What is emphasis and how is it coded. *Pages 748–751 of: Proc. of Speech Prosody*.

Krifka, Manfred. 1999. At least some determiners aren't determiners. *Pages 257–291 of:* Turner, Ken (ed), *The Semantics/Pragmatics Interface from Different Points of View (Current Research in the Semantics/Pragmatics Interface)*, vol. 1. Emerald Group Publishing Limited.

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

Kursat, Leyla, & Degen, Judith. 2020. Probability and processing speed of scalar inferences is context-dependent. *Pages 1236–1242 of:* Denison, Stephanie, Mack, Michael, Xu, Yang, & Armstrong, Blair C. (eds), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

Levinson, Stephen C. 2000. *Presumptive Meanings*. MIT Press Ltd.

Levy, Roger. 2008. Expectation-based syntactic comprehension. *Cognition*, **106**(3), 1126–1177.

Louis, Annie, Roth, Dan, & Radlinski, Filip. 2020. "I'd rather just go to bed": Understanding Indirect Answers. *Pages 7411–7425 of: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics.

Matsumoto, Yo. 1995. The conversational condition on Horn scales. *Linguistics and Philosophy*, **18**(1), 21–60.

McNally, Louise. 2017. Scalar alternatives and scalar inference involving adjectives: A comment on van Tiel, et al. 2016. *Asking the right questions: Essays in honor of Sandra Chung*, 17–28.

Niebuhr, Oliver. 2010. On the Phonetics of Intensifying Emphasis in German. *Phonetica*, **67**(3), 170–198.

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

Papafragou, Anna, & Tantalou, Niki. 2004. Children's Computation of Implicatures. *Language Acquisition*, **12**(1), 71–82.

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

Ronai, Eszter, & Fagen, Lucas. 2022. Exclusives vary in strength and scale structure: experimental evidence. *Pages 258–266 of:* Degano, Marco, Roberts, Tom, Sbardolini, Giorgio, & Schouwstra, Marieke (eds), *Proceedings of the 23rd Amsterdam Colloquium*.

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

Ronai, Eszter, & Xiang, Ming. 2021b. Pragmatic inferences are QUD-sensitive: an experimental study. *Journal of Linguistics*, **57**(4), 841–870.

Ronai, Eszter, & Xiang, Ming. 2022. Three factors in explaining scalar diversity. *Pages 716–733 of:* Gutzmann, Daniel, & Repp, Sophie (eds), *Proceedings of Sinn und Bedeutung*, vol. 26.

Rooth, Mats. 1985. *Association with focus*. Ph.D. thesis, University of Massachusetts, Amherst.

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

Schwarz, Florian, Jérémy, Zehr, Daniel, Grodner, & Bacovcin, Hezekiah Akiva. 2016. *Subliminal Priming of Alternatives Does Not Increase Implicature Responses*. Poster presented at the Logic and Language in Conversation Workshop, University of Utrecht.

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

Sun, Chao, Tian, Ye, & Breheny, Richard. 2018. A Link Between Local Enrichment and Scalar Diversity. *Frontiers in Psychology*, **9**.

Sun, Chao, Tian, Ye, & Breheny, Richard. 2023. A corpus-based examination of scalar diversity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

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

van Tiel, Bob, Van Miltenburg, Emiel, Zevakhina, Natalia, & Geurts, Bart. 2016. Scalar Diversity. *Journal of Semantics*, **33**(1), 137–175.

Ward, Gregory, & Hirschberg, Julia. 1985. Implicating Uncertainty: The Pragmatics of Fall-Rise Intonation. *Language*, **61**(4), 747–776.

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

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

Zondervan, Arjen, Meroni, Luisa, & Gualmini, Andrea. 2008. Experiments on the Role of the Question Under Discussion for Ambiguity Resolution and Implicature Computation in Adults. *Pages 765–777 of:* Friedman, Tova, & Ito, Satoshi (eds), *Proceedings of Semantics and Linguistic Theory (SALT) 18*.

9 Appendix

In recent work, Sun *et al.* (2023) describe the variation across scales in SI calculation (i.e., scalar diversity) in terms of standard deviation, range and variance. They also use Levene's test to compare the equality of variances across different scalar diversity experiments, comparing their own data to van Tiel *et al.* (2016) and Sun *et al.* (2018). In addition to using relative entropy for this purpose in our paper, here we also provide the same measures of variance for our four experiments (Table 3), and conduct further analyses using Levene's test.

Measure	Experiment 1 (SI)	Experiment 2 (weak QUD)	Experiment 2 (strong QUD)	Experiment 3 (*only*)	Experiment 4 (QUD + *only*)
Average	38.71%	34.53%	61.23%	68.42%	88.63%
SD	30.09%	25.03%	23.88%	17.14%	8.1%
Range	2.5%-95%	2.56%-89.75%	5.13%-97.62%	32.5%-100%	67.5%-100%
Variance	905.61	626.4	570.37	293.64	65.56

Table 3: Mean and measures of variance for each experiment, following Sun *et al.* (2023)

We start by comparing the variance in our four experiments to one another. Levene's test shows that the strong QUD manipulation of Experiment 2 had significantly different variance as compared to the baseline SI case of Experiment 1 ($F(1, 118)= 5.69$, $p <0.05$); based on the variance values provided in Table 3, we can observe that this is due to a reduction in variance under the strong QUD. The weak QUD condition, however, did not lead to significantly different variance than Experiment 1 ($F(1, 118)= 3.42$, $p =0.067$). There was a significant difference between the variance of Experiment 3 (*only*) and Experiment 1 ($F(1, 118)= 23.02$, $p <0.001$) as well as Experiment 3 and Experiment 2's strong QUD condition ($F(1, 118)= 6.99$, $p <0.01$), with Experiment 3 having lower variance than either. Lastly, the strong QUD + *only* manipulation (Experiment 4) showed significantly different (reduced) variance as compared to all other experiments: $F(1, 118)= 76.95$, $p <0.001$ with Experiment 1; $F(1, 118)= 57.99$, $p <0.001$ with the strong QUD condition of Experiment 2; $F(1, 118)= 39.69$, $p <0.001$ with Experiment 3.

These results are all in line with those based on the relative entropy measure. Namely, we can see that both a supportive QUD and alternative exclusion via the focus particle *only* lead to reduced scalar diversity, with the effects of the latter being stronger. This conclusion is supported by both the variance values in Table 3 and the F values from Levene's test: the *only* vs. Experiment 1 comparison resulted in a larger F value than the strong QUD vs. Experiment 1 comparison. Lastly, it was the combination of the strong QUD and *only* (Experiment 4) that led to the most different variance from the baseline SI case (i.e., largest F value), which was due to a decrease in variance (i.e., lowest variance value).

Let us now turn to comparing our experimental findings to a uniformity baseline, where all scales lead to the same SI rate. Levene's test shows that the variance of the hypothetical uniform SI rates is significantly different from those of our experiments: $F(1, 117) = 143.27$, $p < 0.001$ with Experiment 1; $F(1, 117) = 98.77$, $p < 0.001$ with the weak QUD from Experiment 2; $F(1, 117) = 142.98$, $p < 0.001$ with the strong QUD from Experiment 2; $F(1, 117) = 163.89$, $p < 0.001$ with Experiment 3; $F(1, 117) = 100.73$, $p < 0.001$ with Experiment 4.

In addition to determining whether variances are significantly different, when comparing different experimental data sets to Experiment 1's SI calculation, we were also able to draw conclusions from the F values regarding which experimental manipulation led to the largest difference in variance. We want to argue, however, that it is less straightforward to derive such an interpretation of F values in the case of comparing experimentally collected data sets to idealized uniformity, i.e., to be able to conclude which experiment is closest to uniformity. This is because the uniformity baseline amounts to zero variance, so any naturally occurring data set will be more varied. The comparison between the two leads to more inflated F values; in our case, the F values for our experimental data compared to uniformity (range of F values: 98.77-163.89) are an order of magnitude higher than when comparing experiments to one another (range: 3.42-76.95). If the intuitive generalization is that very large F values mean very different variances, then the F value of 100.73 for Experiment 4 compared to uniformity would seem to suggest that Experiment 4's results are in fact substantially different from uniformity (note that the F value for the Experiment 1-Experiment 4 comparison was "only" 76.95). Yet, the observation we wish to draw attention to is that Experiment 4 resembles the uniform distribution more than any other experimental data set —an observation which is captured by a reduction in relative entropy. For this reason, if our goal is to compare to uniformity, it is perhaps more informative to simply look at the descriptive measures of variance in Table 3 or to rely on relative entropy.