Three factors in explaining scalar diversity[1]

Eszter RONAI — *The University of Chicago*
Ming XIANG — *The University of Chicago*

Abstract. Rates of scalar inference, whereby the utterance of a weaker term (e.g., *some*) leads hearers to infer the negation of a stronger term (*all*), have been found to vary substantially across lexical scales. For instance, the *some but not all* scalar inference arises much more robustly than *good but not excellent*. This finding has been termed *scalar diversity*. In this paper, we first replicate scalar diversity on 60 different pairs of scalar expressions, which represent a better balance across grammatical categories than has been tested in previous work. We then turn to the open question of what can explain scalar diversity, proposing three factors: 1) a language production-based metric of how accessible the stronger alternative (*all*) is; 2) the distinctness of the two scalar terms (*some* vs. *all*), as measured by posterior degree estimates; and 3) the meaning of the negated strong scalar term (*not all*), again measured by degree estimates. We report on three experiments showing that these factors can indeed explain some of the observed variation in scalar inference rates.

Keywords: experimental pragmatics, scalar inference, scalar diversity.

1. Background

In communication, the utterance of a weaker proposition can lead hearers to infer the negation of a stronger alternative proposition. This process gives rise to scalar inference (SI), whereby, for instance, utterances including *some* have the pragmatic meaning *some but not all*. This is exemplified in (1).

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

A standard (Neo-)Gricean account of the inferential process underlying SI calculation can be summarized as follows. Hearers assume that speakers are trying to be maximally informative while remaining truthful. Lexical items such as *some* and *all* form an informativity scale on which *all* is informationally stronger, i.e. more informative, than *some*. When someone hears the utterance in (1), she reasons that, if the stronger alternative *Mary ate all of the cookies* were true, the speaker would have said that. Because the speaker chose not to say it, the hearer can infer its negation, arriving at the SI-enriched meaning in (1b), rather than the literal meaning in (1a) (Grice, 1967; Horn, 1972).

A large amount of experimental research has concentrated on the <*some, all*> (and to a lesser extent, the <*or, and*>) scale; see van Tiel et al. (2016: p.139, Table 2) for a summary. But there

[1]We would like to thank Chris Kennedy, Hannah Rohde, Michael Tabatowski, and the audiences at Sinn und Bedeutung 26 and at the University of Pennsylvania Language & Cognition Lab for helpful discussion and feedback. We are also grateful to Sherry Yong Chen for technical help with some experimental tasks. This material is based upon work supported by the National Science Foundation under Grant No. #BCS-2041312. All mistakes and shortcomings are our own.

exist many other pairs of lexical items that form a scale and can give rise to SI. The example in (2), for instance, demonstrates SI on the <*good, excellent*> scale.

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

Upon encountering the utterance in (2), hearers may reason about the stronger alternative *The movie is excellent*, and based on its negation compute the SI in (2b), going beyond the literal meaning in (2a). This parallels the SI calculation in (1). However, experimental investigations of different lexical scales have revealed that they actually vary considerably in how likely they are to lead to SI calculation: the *some but not all* SI, for instance, arises more robustly than the *good but not excellent* SI. This variation across scales has been termed *scalar diversity*. The first large-scale study on scalar diversity was conducted by van Tiel et al. (2016), who tested 43 different lexical scales and found rates of SI calculation ranging from 4% (for seven scales) to 100% (for two scales); see also Baker et al. (2009); Doran et al. (2012); Beltrama and Xiang (2013) for earlier findings.

A prominent research question regarding scalar diversity is: what properties of scales predict the likelihood of scalar inference? Existing work has identified a number of such properties. Van Tiel et al. (2016) put forth two hypothesized factors that might explain scalar diversity: the availability and distinctness of lexical scales. Availability is relevant because, in order for SI to arise, hearers must assume that the speaker considered using a stronger alternative (e.g., *all*) to what she ultimately uttered (*some*); in other words, they must assume that the stronger scalar term was available to the speaker. Distinctness refers to whether the speaker considered the distinction between the weaker (*some*) and stronger (*all*) scalar terms substantial enough that she would have used the stronger one if possible. Of the two factors, distinctness was found to be a predictor of SI rates. Van Tiel et al. (2016) operationalized distinctness as semantic distance and boundedness, two notions that go back to Horn (1972: p. 112). Measuring semantic distance via a rating task, it was revealed that the more distant a weak and a strong scalar term are, the stronger the SI from the weak term is. This can be intuitively seen on the <*some, many, most, all*> scale: an utterance of *Mary ate some of the cookies* most strongly implicates that Mary didn't eat all of the cookies, while the inference *Mary didn't eat most of the cookies* is less likely, and *Mary didn't eat many of the cookies* least likely. The second component of distinctness is boundedness: unbounded scales (e.g., <*good, excellent*>), in which both the weaker and stronger term denote intervals, were found to lead to significantly fewer SIs than bounded scales, in which the stronger scalar term denotes a fixed point or endpoint (e.g., <*some, all*>). As measures of availability, van Tiel et al. (2016) considered association strength between the weaker and stronger scalar terms, frequency, grammatical class, and the two scalar terms' semantic relatedness —but none of these was found to be a significant predictor of SI rates in the authors' experiments. In later work, however, Westera and Boleda (2020) showed that a sufficiently fine-grained notion of semantic relatedness (derived from distributional semantics) does predict SI rates across scales, but there is a negative correlation: the more semantically similar two scalar terms are, the lower the SI rate. The authors argue that this is because semantic relatedness in fact indexes distinctness: the more similar two terms are, the less distinct they are, and hence the lower the likelihood of SI.

Three factors in explaining scalar diversity

Subsequent work has identified further properties of scales that predict how likely they are to lead to SI. Investigating adjectival scales, Gotzner et al. (2018) found that certain semantic properties of adjectives, such as polarity and extremeness, are relevant for SI calculation. In particular, their results revealed that negative scales (e.g., *<bad, awful>*) yield higher SI rates than positive ones (e.g., *<good, great>*). Additionally, scales in which the stronger term is an extreme adjective (e.g., *excellent* or *huge*) were found to lead to lower SI rates —for findings regarding extremeness, see also Beltrama and Xiang (2013). Existing work has also related scalar diversity to other semantic-pragmatic processes. Sun et al. (2018) investigated propensity for local enrichment, indexed by the naturalness of sentences such as *Mary ate all, so not some, of the cookies*. This factor was positively correlated with SI rates: as the authors argue, in order for a sentence such as *Mary ate all, so not some, of the cookies* to be natural and not contradictory, *some* has to be locally be interpreted on its SI-enriched meaning (*some but not all*). Lastly, Gotzner et al. (2018) also showed that SI rates are negatively correlated with the degree of negative strengthening of the stronger scalar term. Negative strengthening is the phenomenon whereby *John is not brilliant* is interpreted as conveying that John is rather stupid, which can be analyzed as a manner implicature (Horn, 1989). In Gotzner et al. (2018)'s study, participants saw sentences such as *He is not brilliant*, and were asked whether they can conclude from this statement *He is not intelligent*. Endorsements of this conclusion were negatively correlated with SI rates, suggesting that, at least for some scales, scalar and manner implicatures might stand in competition (Levinson, 2000).

The observed variation in SI rates has been also related to properties of the context, broadly construed. Pankratz and van Tiel (2021) offer a usage-based explanation of scalar diversity, and show that it is predicted by the relevance of the SI at hand. Specifically, they 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., *It's good but not excellent*) in a corpus. Ronai and Xiang (2021) investigated the role of the Question Under Discussion in explaining scalar diversity. They hypothesized that, given that experiments typically present SI-triggering sentences in the absence of any context, variation across scales in what implicit discourse context they bring to mind affects their likelihood of leading to SI calculation. Their study indeed found that the more likely people are to ask a polar question involving the stronger scalar (e.g., *Is the movie excellent?*), the higher the rate of SI calculation, but with the caveat that this correlation only holds for bounded scales.

Though existing work has identified some significant predictors of scalar diversity, a substantial amount of the statistical variance in SI rates is still unaccounted for. Van Tiel et al.'s (2016) statistical analysis revealed that semantic distance explained 10% of the observed variance, while boundedness explained only 3%. Sun et al. (2018) found that 15% of the variance was explained by propensity for local enrichment. In Gotzner et al.'s (2018) study, extremeness captured 17% and polarity 5% of the variance. While Westera and Boleda's (2020) results did reveal an effect of semantic relatedness (contra van Tiel et al. 2016), this metric still only captured 4-6% of the variance. Lastly, Pankratz and van Tiel (2021) found that relevance explained 4%. Models that combine multiple known predictors from existing studies still fall short of explaining all of the observed variance: Sun et al.'s (2018) best fitted model explained 63%, Gotzner et al.'s (2018) 62%, while Pankratz and van Tiel (2021) report that their model combining relevance with other predictors explained 8%. This suggests that a lot of scalar

diversity is still unexplained, and thus in the present paper we propose three factors that further predict SI rates across scales.

The remainder of this paper is structured as follows. We first report on a corpus study conducted to collect a set of lexical scales (Section 2), and then we replicate scalar diversity (Experiment 1, Section 3). The three factors we test as explanations of scalar diversity are the accessibility of the stronger alternative (Experiment 2, Section 4), the distinctness of scalar terms (Experiment 3, Section 5), and the meaning of the negated stronger alternative (Experiment 2, Section 6). Section 7 analyzes the variance explained by these, and Section 8 concludes.

2. Corpus study

In previous work on scalar diversity, the set of scales studied included mostly (70%, e.g., van Tiel et al. 2016; Sun et al. 2018) or entirely (e.g., Gotzner et al. 2018; Pankratz and van Tiel 2021) adjectival scales. But if our goal is to identify properties of SI that hold generally, across all scales, then we should also investigate scales from other grammatical classes. For this reason, we created a new set of scales by taking existing sets from van Tiel et al. (2016) and de Marneffe and Tonhauser (2019) and supplementing them with corpus work. Specifically, we conducted the following corpus searches in the Corpus of Contemporary American English (Davies, 2008): *X or even Y*; *not just X but Y*; *X but not Y*. These searches were conducted for adjectives, verbs, and adverbs. The expectation is that these 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 in which *X* and *Y* were clearly not in a scale-mate relation were discarded. Combining the items from the two published studies with the corpus data resulted in a total of 101 items.

In the next step, the following semantic tests were used to filter the results, probing whether *X* and *Y* indeed form a scale. The question in (3a) tests 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 tests in (3b)-(3c) probe for asymmetric entailment (Horn, 1972): *Y* should entail *X*, but not vice versa, in order for *X* and *Y* to qualify as scale-mates.

(3) a. Is *X and even Y* odd? Expected answer: No
b. Is *X but not Y* contradictory? Expected answer: No
c. Is *Y but not X* contradictory? Expected answer: Yes

Wherever a pair did not produce the "expected answer", it was excluded. Lastly, wherever a lexical item participated in more than one scale, one of those scales was excluded, e.g., because *good* occurred in both <*adequate, good*> and <*good, excellent*>, the former scale was excluded. The resulting final set consists of 60 lexical scales, which form the basis of all experiments reported in this paper.

3. Experiment 1: Scalar diversity

Experiment 1 adapted the inference task of van Tiel et al. (2016) to test the rate of SI calculation, and replicate the scalar diversity effect, on our expanded set of 60 different lexical scales.

3.1. Participants, task and procedure

42 native speakers of American English participated in an online experiment, administered on the Ibex platform (Drummond, 2007). Participants were recruited on Prolific and compensated $2. Native speaker status was established via a language background survey, where payment was not conditioned on participants' responses. Data from 40 participants is reported below.

An inference task was used to investigate the likelihood of deriving an SI from 60 different scales. 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. A "Yes" answer indicates that the participant has calculated the relevant SI (*good but not excellent*), while a "No" answer indicates that the participant has not calculated the SI, i.e. they are interpreting *good* as meaning *good and possibly excellent*.

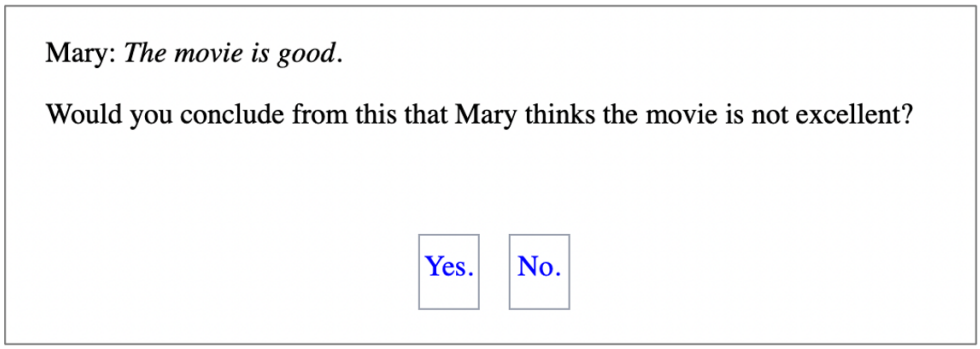

Figure 1: Example experimental trial from Experiment 1

7 filler items were also included, which contained two terms that are either in an entailment relation (*wide* → *not narrow*), or unrelated (*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.

3.2. Prediction

Existing literature has consistently found robust variation across scales in how likely they are to lead to SI calculation. Given this, we predict that our Experiment 1 will replicate the finding of scalar diversity. That is, the percentage of "Yes" vs. "No" responses in the inference task is predicted to vary substantially across the 60 different scales.

3.3. Results and discussion

Figure 2 shows the results of Experiment 1, where the percent of SI calculation corresponds to the proportion of "Yes" responses averaged across participants. As can be seen in the figure,

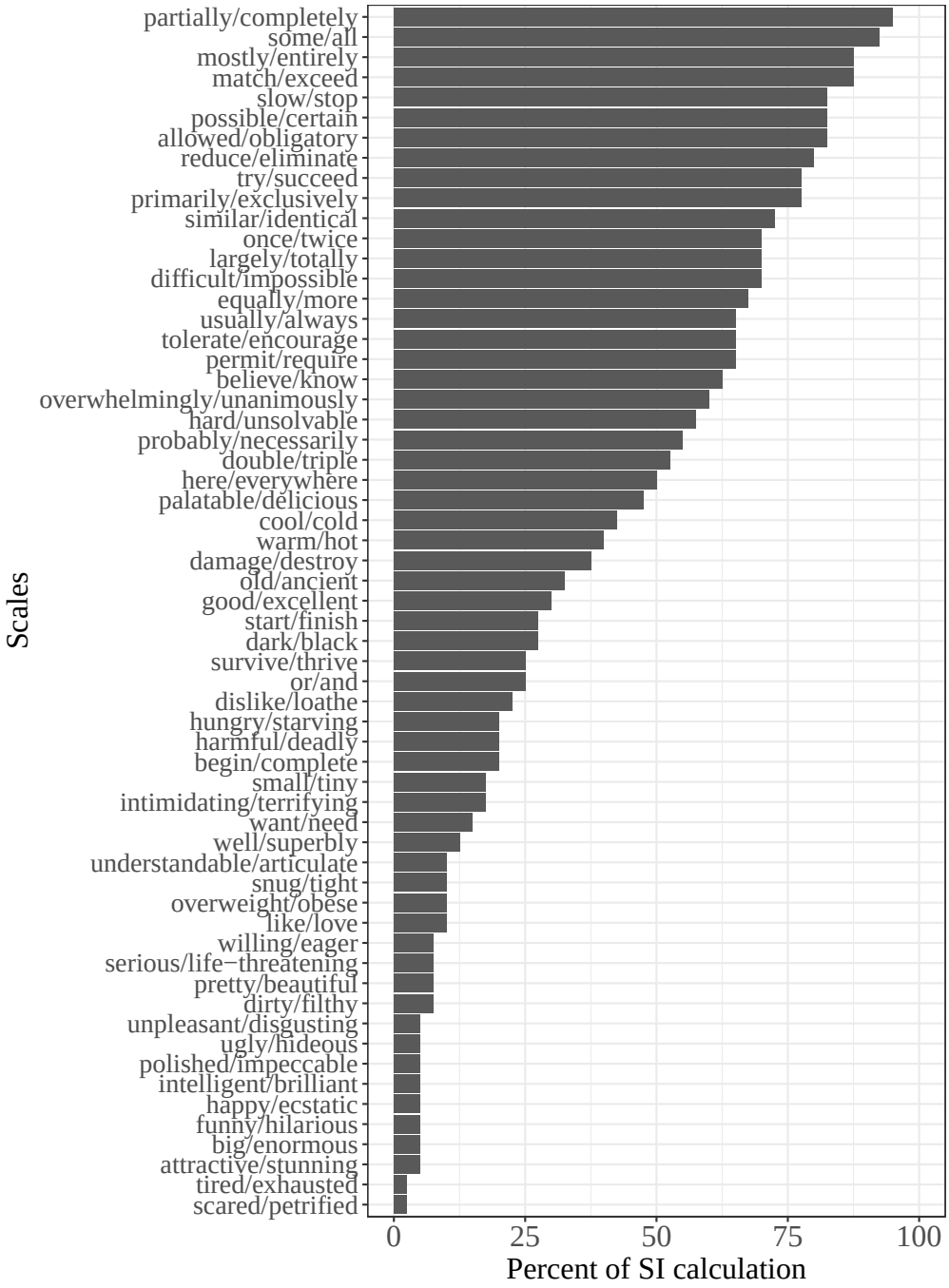

Figure 2: Results of Experiment 1: SI rate for 60 different scales

Experiment 1 revealed considerable variation across scales, with the rate of SI calculation ranging along a continuum from 2.5% (for <*scared, petrified*> and <*tired, exhausted*>) to 95% (for <*partially, completely*>). Experiment 1 thus successfully replicates the scalar diversity effect on a set of 60 different scalar expressions that represent a better balance of grammatical categories than was tested in previous work.

4. Experiment 2: Accessibility of stronger alternative

In Experiment 2, we used a modified cloze task as a metric for the accessibility of stronger alternatives across scales, which was found to significantly predict the rate of SI calculation from Experiment 1.

4.1. Hypothesis

We hypothesize that scalar diversity can, in part, be explained by how accessible a stronger alternative is, given the weaker scalar. The causal mechanism behind this hypothesis is as follows. We assume that SI calculation proceeds via reasoning about alternatives, and that hearers generate a set of alternatives when they encounter a potentially SI-triggering utterance that contains a weak scalar term. The more accessible an alternative is, the more likely hearers are to reason about it, and therefore the more likely the relevant SI is to arise. In the context of scalar diversity, the intuition is that there may be differences across scales in how strongly the weaker scalar evokes a stronger alternative. For instance, it is possible that when encountering a sentence containing *some*, the stronger alternative *all* always comes to mind; but when encountering a sentence containing *good*, a number of competing alternatives may be activated, such as *excellent, funny, thrilling, thought-provoking*, and so on.

4.2. Participants, task and procedure

61 native speakers of American English participated in an online (Ibex) experiment for $2 compensation. Participant recruitment and screening was identical to Experiment 1. Data from all 61 participants is reported below.

We operationalize alternative accessibility as cloze probability, a commonly used measure of the predictions the parser makes in language comprehension. In particular, the probability of a target word completing a given sentence frame is taken to index how expected a word is in a context (Taylor 1953; see also i.a. Kutas and Hillyard 1984). Our experiment employed a modified cloze task: participants were presented with a dialogue context where Sue uttered a potentially SI-triggering sentence, such as *The movie is good* (identical Experiment 1), and Mary followed up by saying *So you mean it's not BLANK*. Participants were instructed to complete Mary's utterance with the first word that came to mind, making sure that their completion made sense in the context of the dialogue. Figure 3 shows an example trial item.

Similarly to Experiment 1, Experiment 2 included 60 experimental trials and 7 fillers to serve as catch trials. The 2 practice trials at the beginning of the experiment also provided participants with some feedback on what is a reasonable completion in the cloze task. Experiment 2 included two within-participants conditions that addressed different research questions and are not discussed here; due to counterbalancing, we therefore ended up collecting 19-22 completions per scale. The experiment was administered in a Latin Square design.

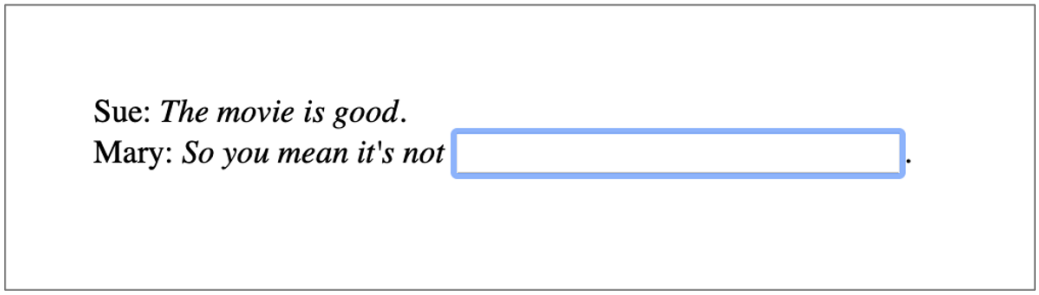

Figure 3: Example experimental trial from Experiment 2

4.3. Prediction

Given the accessibility hypothesis, the prediction we make for the results of Experiment 2 is that the more frequently the stronger alternative (e.g., *all*) is mentioned in the modified cloze task, the higher the SI rate for that scale (*some but not all*) from Experiment 1.

4.4. Results and discussion

The results of the cloze task were coded as follows. We counted the number of times the relevant alternative was mentioned, and divided this by the number of total completions for that scale, resulting in the percent of stronger alternative mentioned. Figure 4 shows these results, correlated with SI rates from Experiment 1. In the coding of the results, synonyms of the stronger alternative were also counted. There was a positive correlation between the results of Experiment 2 and 1 (Pearson's correlation test: r=0.59, $p<0.001$). The higher the percent of mentioning the stronger alternative (*excellent*) in the cloze task, the higher the corresponding SI rate from that scale (*good but not excellent*). For a more detailed statistical analysis, combining all of our experiments and analyzing SI calculation as "Yes" vs. "No" responses, see Section 7.

In other words, scalar diversity was shown to be predicted by the accessibility of the stronger scalar —that is, by how strongly a weaker scalar evokes a stronger alternative. To provide a few illustrative examples beyond the overall quantitative analysis, for some scales in Experiment 2, the stronger alternative was given almost every time as a cloze completion; for instance, for *some*, the alternative *all* was provided by almost all participants (with one participant providing *most*). On the opposite end, for some scales the stronger alternative from Experiment 1's inference task was never provided: for instance, for *good*, the completions included *bad, terrible, overrated*, but not *excellent*. This suggests that the relevant stronger alternative is not very accessible here. Impressionistically, in such cases, antonyms to the weaker scalar term were frequent completions. Lastly, some scales led to a greater variety of completions, suggesting that a larger number of (not just scalar) alternatives can be activated upon encountering the SI-triggering utterance: for *try*, for example, participants filled in the stronger alternative *succeed*, but also *fail, surrender, concede, quit*.

A potential caveat to mention is that our measure of accessibility may be interpreted as the production side of scalar diversity. In the inference task of Experiment 1, participants have

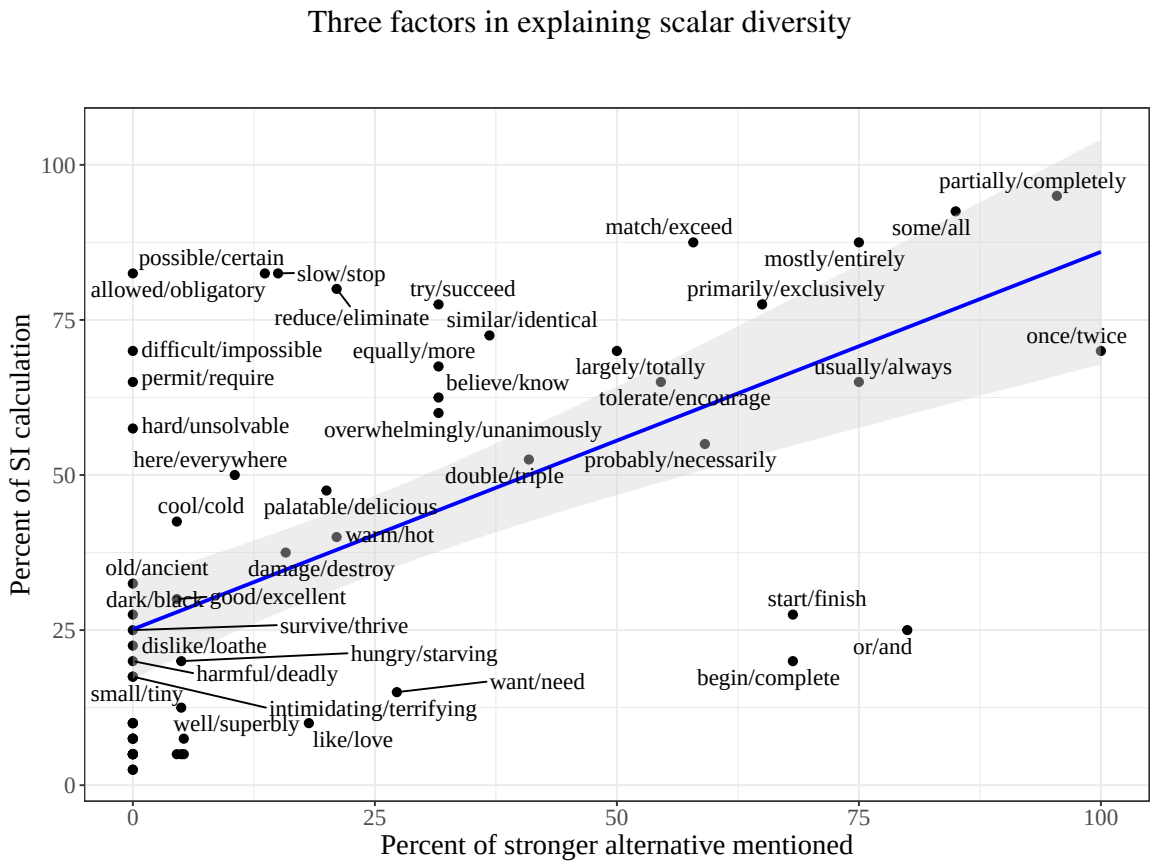

Figure 4: Results of Experiments 1 and 2: The *x* axis shows alternative accessibility from Experiment 2. The *y* axis shows SI rate from Experiment 1.

to judge statements containing the negated stronger scalar (*not excellent*), having seen an SI-triggering sentence. In Experiment 2, we asked participants to fill in a blank under negation (*So you mean it's not...*), as a response to the same SI-triggering sentences. Perhaps the reason that the results of the cloze task predict SI rates, and both experiments show diversity, is that we are tapping into outcomes of the same mechanism, with Experiment 1 testing the comprehension and Experiment 2 the production side. However, there is one important difference we would like to highlight: the inference task asks participants to make a decision about a particular stronger alternative. The cloze task, on the other hand, probes what is a relevant contrast for a weaker scalar term—e.g., is it *good* vs. *excellent*, or *good* vs. *bad*. Ultimately, the cloze task is therefore informative regarding whether the specific alternative message that the hearer infers the negation of – having seen an SI-triggering utterance – is the same as the stronger alternative from the lexical scale that we test in the inference task.

Our proposal of alternative accessibility is closely related to van Tiel et al.'s (2016)'s proposal that the availability of the stronger alternative should predict scalar diversity. The authors argue that for SI to arise, it has to be the case that the speaker could have actually considered using the stronger scalar term instead of the weaker one she uttered. As mentioned in Section 1, van Tiel et al. (2016) tested four different operationalizations of availability, but none of them were found to be a predictor of diversity. Our operationalization is novel in that it utilizes a language production task in a discourse context, and it does end up predicting SI rates.

5. Experiment 3: Distinctness of scalar terms

In Experiment 3, we used posterior degree estimates as a metric for the distinctness of the weak and strong scalar terms, which significantly predicted the likelihood of SI calculation.

5.1. Hypothesis

Distinctness of scalar terms was originally put forth by van Tiel et al. (2016) as a potential explanation for scalar diversity. Distinctness is relevant for the likelihood of SI calculation for the following reason. The inferential process underlying SI calculation involves the hearer reasoning about, and negating, a stronger alternative (*all*) that the speaker could have said, but did not. For this reasoning to go through, there has to be a clear stronger alternative, and it has to be sufficiently stronger. In other words, the more distinct two scalar terms (*some* vs. *all*) are, the more likely the hearer is to assume that the speaker should have used the stronger term if possible. If it is difficult to distinguish the weak and strong scalar, e.g. if they are near-synonyms, SI calculation is unlikely.

5.2. Participants, task and procedure

60 native speakers of American English participated in an online (Ibex) experiment for $2 compensation. Participant recruitment and screening was identical to Experiment 1. Data from all 60 participants is reported below.

Our operationalization for distinctness of scalar terms is inspired by Bayesian pragmatics, which assumes and models recursive reasoning between speaker and hearer (i.a. Goodman and Frank, 2016; Lassiter and Goodman, 2017). In Experiment 3, we are interested in what world states hearers think utterances such as *The movie is good* vs. *The movie is excellent* describe. To determine this, we experimentally collect degree estimates on the underlying scales. In other words, what we are testing is: after encountering the relevant utterance, what degree of goodness do hearers ascribe to the movie?

Experiment 3 therefore employed a degree estimate task. Participants were presented with a sentence such as *The movie is good* or *The movie is excellent*, and were instructed to answer a question like *On a 0-100 scale, how good is the movie?* by picking a point on a scale from 0 to 100. The weak and strong scalar terms were tested as a between-participants manipulations (30 participants in each condition). Figure 5 shows an example trial item from the strong scalar term condition. We aimed to create neutral questions that would not bias participants toward either end of the scale. For adjectival lexical scales, questions relied on the weaker term wherever possible (*On a 0-100 scale, how old is the house?* for <*old, ancient*>), while in other cases we picked a neutral underlying adjective, e.g., *On a 0-100 scale, how likely is success?* for <*possible, certain*>. Questions for verbal and adverbial scales were necessarily more varied, but aimed to be neutral and refer to the underlying scale, e.g., *On a 0-100 scale, how much will the sales increase?* for <*double, triple*> or *On a 0-100 scale, how often is the lawyer early?* for <*usually, always*>. This task is an idealization, because not all lexical scales

map onto a bounded underlying degree scale, but results suggest that participants were able to accommodate and make sense of the task in the context of this experiment.

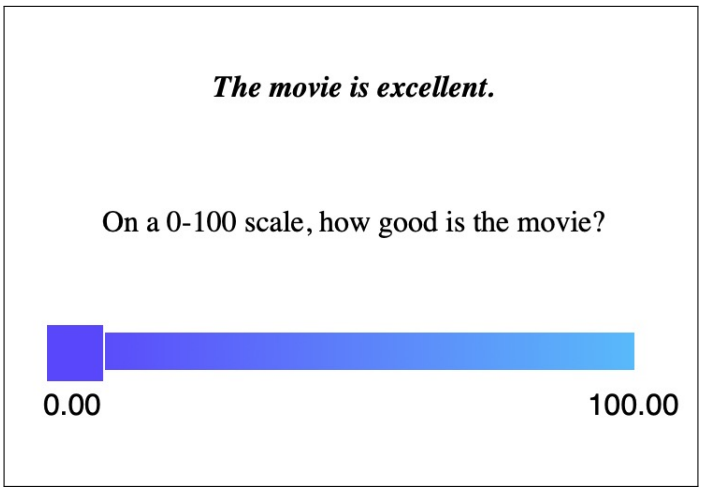

Figure 5: Example experimental trial from Experiment 3: stronger scalar term

The experiment included 60 critical items, as well as 3 practice trials and 5 filler items. The latter served as catch trials and used antonyms in the sentence and task question, e.g., *The table is clean* was paired with *On a 0-100 scale, how dirty is the table?*.

5.3. Prediction

The data collected in Experiment 3 represents hearers' probabilistic guesses on what world state the speaker has in mind, given her utterance. Based on the distinctness hypothesis, we predict that the greater the difference between the degree estimates for the weak and the strong scalar terms, i.e. the further apart they are on the underlying degree scale, the higher the SI rate will be for that scale. As mentioned, this is because for an SI (*good but not excellent*) to arise, *good* and *excellent* have to be perceived as describing two different world states.

5.4. Results and discussion

Averaged over the 60 lexical scales, the stronger scalar terms received higher ratings than the weaker terms —see Figure 6. In other words, a sentence such as *The movie is excellent* led hearers to attribute a higher degree of goodness to the movie than *The movie is good*. This difference is statistically significant: using the lme4 package in R (Bates et al., 2015), we fit a linear mixed effects regression model that predicted Response (0-100) by Condition ("weak" and "strong", as well as "not strong" from Experiment 4). The fixed effects predictor Condition was treatment-coded, with weak as the reference level. Random intercepts were included for participants and items. Responses to strong terms were found to be significantly higher than to weak terms (Estimate=22.68, Std. Error: 2.68, t=8.38, $p<0.001$). This serves as confirmation that participants were performing the task adequately.

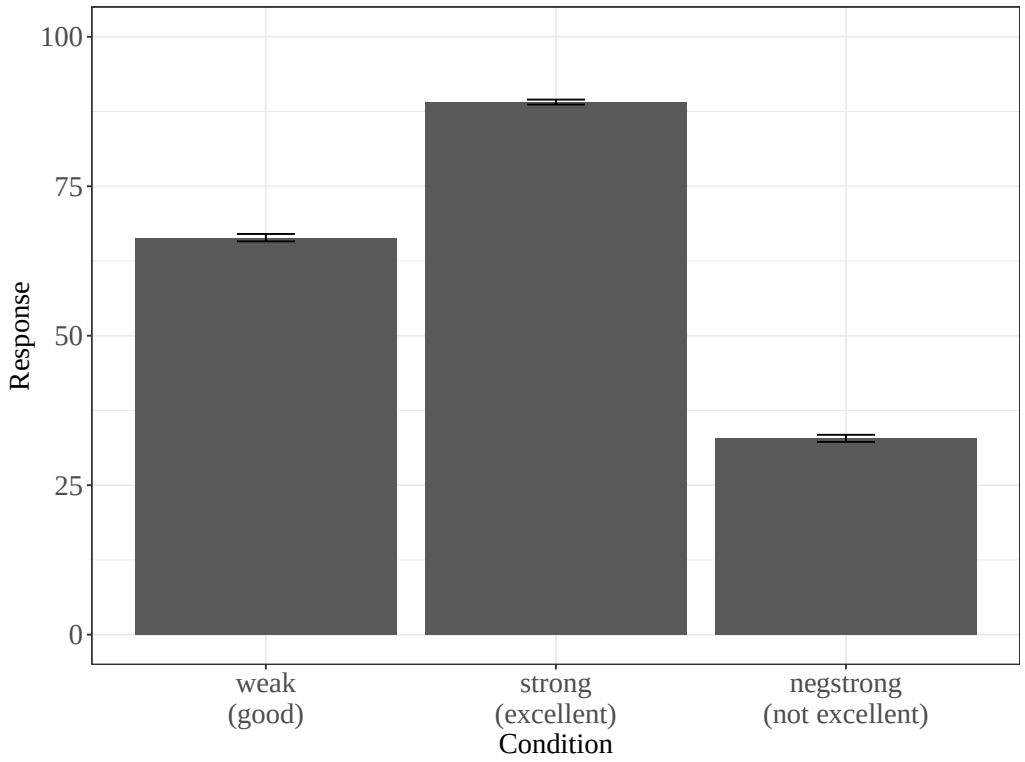

Figure 6: Results of Experiment 3 and 4: Average response on the degree estimate task (0-100 scale) for the three different conditions

To check the prediction of distinctness, we took the absolute difference in means between the weak and strong terms: for instance, *The movie is good* received a response of 69.4 on the 0-100 scale, while *The movie is excellent* received 89.1, resulting in a "distinctness" value of 19.7. Figure 7 shows these results, correlated with SI rates from Experiment 1. As can be seen in the figure, there was a positive correlation between the results of Experiment 3 and 1 (Pearson's correlation test: r=0.33, $p<0.05$). That is, scalar diversity was shown to be predicted by the distinctness of scalemates. Specifically, the higher the difference between a weak (*good*) and a strong (*excellent*) term, as measured via degree estimates, the higher the corresponding SI rate from that scale (*good but not excellent*).

In other words, we found that the more distinct the world states that the weaker and the stronger term are taken to describe, the higher the SI rate for that scale. Experiment 3's results thus present further evidence for van Tiel et al.'s (2016) distinctness hypothesis, using a novel operationalization that relies on empirically collected posterior degree estimates. Van Tiel et al. relied on the notion of boundedness, as well as experimentally collected judgements about semantic distance, to test the distinctness hypothesis. It is worth discussing how the latter relates to our Experiment 3. In the semantic distance experiment, participants were presented with a pair of sentences, such as *She is intelligent* and *She is brilliant*. They then had to respond to the question *Is statement 2 stronger than statement 1?* via a 7-point Likert scale, where 1 corresponded to "equally strong" and 7 to "much stronger". In line with the distinctness hypothesis, the authors found that semantic distance was positively correlated with SI rates: the more distant a weak and a strong scalar term were in their experiment, the more likely the corresponding

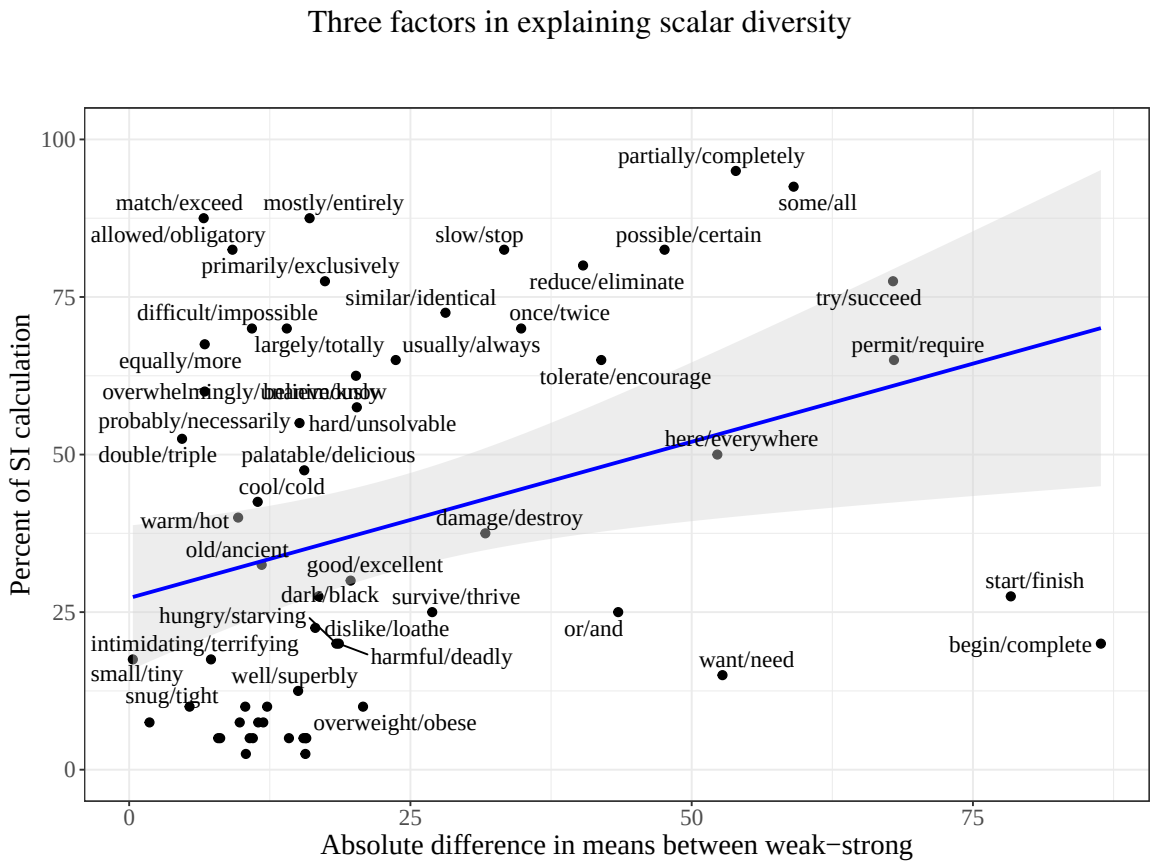

Figure 7: Results of Experiments 1 and 3: The *x* axis shows distinctness between each weak-strong scalar pair from Experiment 3. The *y* axis shows SI rate from Experiment 1.

SI. Our experiment 3 differs from van Tiel et al.'s in that it does not assume any *a priori* strength relation; our experimental instructions did not presuppose that one statement could be stronger than the other, and participants simply picked points on an underlying scale. Another notable difference is that judging the relative strength of statements requires a metalinguistic judgment, and therefore degree estimates are arguably a more natural task. Altogether, the experiment reported here constitutes further evidence for van Tiel et al.'s distinctness hypothesis, going beyond existing evidence in the prior literature.

6. Experiment 4: Meaning of the negated strong scalar

In Experiment 4, we show that scalar diversity is (partially) explained by the meaning of the negated strong scalar term, as compared to the weak scalar. As in Experiment 3, our measure relies on experimentally collected degree estimates.

6.1. Hypothesis

In the previous two experiments, in line with much of the literature on scalar diversity, we focused on potential explanations for scalar diversity that had to do with the relationship between the weak and the strong scalar term. Experiment 4 takes a slightly different perspective, as it

focuses on the meaning of the negated strong term (*not excellent*) as a predictor of the variation in SI rates. Let us first consider the inference task commonly used to test SI calculation, which we also used in Experiment 1. The inference task presents participants with an SI-triggering statement, such as *The movie is good*, and then poses the question: *Would you conclude from this that Mary thinks the movie is not excellent?*. (Neo)-Gricean accounts of SI calculation assume that hearers reason (only) about potential stronger alternatives to the weaker utterance they heard. But given the particulars of the inference task, it is conceivable that the meaning of the *negated* alternative (e.g., *not excellent*) also plays a role. In Experiment 4, we therefore probe what such negated stronger alternative statements mean, and what hearers therefore have in mind when answering the question of the inference task.

The specific hypothesis that we test is that the more similar the weak (*good*) and the negated strong (*not excellent*) term are, i.e. the smaller the difference between them on a degree scale, the higher the SI rate should be for that scale. Suppose, for instance, that *good* and *not excellent* are interpreted as describing two very different world states—that is, they are distant on the degree scale of goodness. In this case, it is implausible for a participant to conclude that a speaker meant *not excellent* when she uttered *intelligent*. This can lead to a low rate of "Yes" responses in the inference task, which is then interpreted as a low SI rate.

6.2. Participants, task and procedure

31 native speakers of American English participated in an online (Ibex) experiment for $2 compensation. Participant recruitment and screening was identical to Experiment 1. Data from all 31 participants is reported below.

Experiment 4 had the same task and procedure as Experiment 3 —see Figure 5 for an example trial item. Here, we tested the negated strong term (in a between-participants design with Experiment 3). That is, participants saw the sentence *The movie is not excellent*, and then had to indicate on a 0-100 scale how good they thought the movie was.

6.3. Prediction

Our prediction for the results of Experiment 4 is that the smaller the difference between the degree estimates for the weak and the negated strong term, the higher the corresponding SI rate will be. In other words, we predict a negative correlation between the weak-not strong difference and SI rates.

6.4. Results and discussion

We conducted the same analyses as those reported in Experiment 3. Responses to negated strong terms were found to be significantly lower than to weak terms (Estimate=-33.59, Std. Error: 2.65, t=-12.65, $p<0.001$) —see Figure 6. That is, sentences such as *The movie is not excellent* received, on average, lower ratings on a 0-100 goodness scale than sentences such as

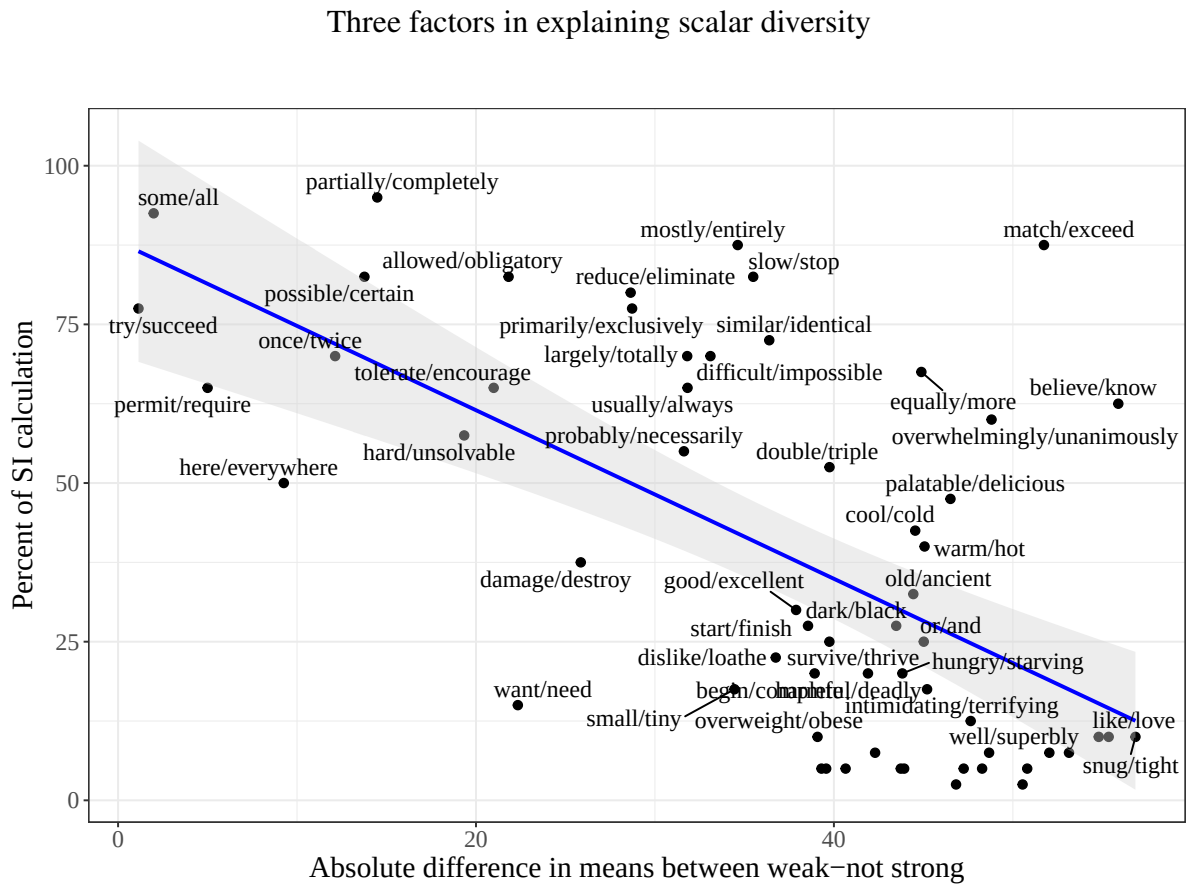

Figure 8: Results of Experiments 1 and 4: The *x* axis shows the meaning of the negated stronger term from Experiment 4. The *y* axis shows SI rate from Experiment 1.

The movie is good. We return to this finding below, in our discussion of negative strengthening. To check our main prediction that the meaning of the negated strong term captures scalar diversity, we again calculated the absolute difference in means between the response to the weak term (Experiment 3) and the response to the negated strong term (Experiment 4). For example, for the <*good, excellent*> scale, *The movie is good* received a response of 69.4 on the 0-100 scale, while *The movie is not excellent* received 31.5, resulting in a score of 37.9 —these are plotted on the *x* axis of Figure 8. There was a negative correlation between the results of Experiment 4 and 1 (Pearson's correlation test: r=-0.61, $p<0.001$); the more similar the world states that a weaker and negated stronger term are taken to describe, the higher the SI rate is for that scale. This constitutes evidence that the meaning of the negated stronger scalar plays a role in scalar diversity.

To reiterate, the motivation for Experiment 4 was that the inference task commonly used to test SI calculation explicitly mentions the negated stronger term, raising the possibility that when participants choose not to endorse the conclusion that a speaker meant *not excellent* by uttering *good*, they do so because they perceive *not excellent* and *good* as meaning different things. Our findings suggest that the meaning of the negated strong term, as measured by experimentally collected degree estimates, indeed captures some of the variation in SI rates that is observed across scales. This raises broader questions about whether the inference task is a good way to measure SI calculation. One limitation of the inference task in its current form is that it explicitly mentions, and therefore makes salient, the stronger alternative to a weaker scalar term

(*Would you conclude... not **excellent**?*). Yet, scalar diversity emerges despite this potentially biasing nature of the task: we do not find that inference rates are uniformly high across scales, even though the stronger alternative is mentioned in the task question. The findings of our Experiment 4 highlight a second potential problem with the inference task: namely, that it might introduce complications not only because it mentions stronger alternatives like *excellent*, but because it mentions *not excellent*, whose meaning we have shown to matter for SI calculation. For more recent discussion about task effects in experimental investigations of SI, see also Sun and Breheny (2021), who found that a task question like *Would you conclude from this that, according to Mary, not all of the questions are easy?* (similar to our Experiment 1) vs. one like *Would you conclude that, it could be that Mary thinks, all of the questions are easy?* produce different results.

As is reflected in the averages reported in Figure 6, the negated strong degree estimate was lower than the weak degree estimate for many lexical scales. This finding can be interpreted as negative strengthening, the pragmatic phenomenon where hearers take *John is not brilliant* to mean not only that John is less than brilliant (the sentence's literal meaning), but that he is less than intelligent, or that he is in fact stupid (Horn, 1989). As discussed in Section 1, Gotzner et al. (2018) experimentally tested propensity for negative strengthening across different scales: participants saw sentences such as *He is not brilliant* and were asked whether they can conclude that he is not intelligent. The authors found that "Yes" responses negatively correlated with SI rates and were able to predict scalar diversity. While negative strengthening is certainly relevant to the results of our Experiment 4, there are a number of important respects in which our findings differ from Gotzner et al.'s. First, our collected data include scales that did not show negative strengthening, i.e. where the negated strong scalar term had a higher rating on the 0-100 degree scale scale than the weak scalar, suggesting that not all of our Experiment 4 findings are attributable to negative strengthening. Second, though arguably tapping into similar pragmatic phenomena, negative strengthening is chiefly about *not brilliant* being interpreted as *not (even) intelligent*, while what we measured in this experiment is whether *not brilliant* is similar to *intelligent* in what world state it is taken to describe.

7. Combined analysis and variance explained

Having seen evidence that our three identified factors (accessibility, distinctness, meaning of the negated strong term) are correlated with SI rates, let us now turn to how much of the observed variance they are able to capture. To test this, we conducted an analysis combining data from all experiments. Using the lme4 package in R (Bates et al., 2015), we fit a logistic mixed effects regression model that predicted Response ("Yes" vs. "No") in Experiment 1's inference task as a function of Accessibility, Distinctness, and Meaning of the negated strong term (Experiments 2-4). The model included random intercepts for participants. The model's estimates are shown in Table 1.

To check how much of the variance in the data is explained, we used the rsq package in R (Zhang, 2021) to compute R^2 values. We found that the model combining all three factors explained 25.8% of the variance in the data, with 22.4% coming from the fixed effects and 3.4% from the random effects. To test what proportion of the variance each factor explains,

Three factors in explaining scalar diversity

	Estimate	Std. Error	z value	p value
Intercept	2.11	0.24	8.74	
Accessibility	0.03	0	13.76	<0.001
Distinctness	-0.03	0	-7.69	<0.001
Meaning of negated strong term	-0.07	0	-15	<0.001

Table 1: Parameter estimates, standard errors, z values and p values from a logistic mixed effects regression model of the "Yes" vs. "No" responses in Experiment 1, predicted by the factors identified in Experiments 2-4

we checked how much the R^2 is reduced by fitting a model that removes that factor. That is, to calculate how much of the variance is explained by Accessibility, we fit a regression model only including the other two factors, and checked how the R^2 of that model compares to 22.4%. (Here, we concentrate only on the variance explained by fixed effects.) Using this method, we found that the accessibility of stronger scalar explains 7.9% of the variance, distinctness between the weak and strong scalar terms constitutes only 2.7% of the variance, while the meaning of the negated strong scalar is the best predictor, capturing 9.4% of the variance.

Altogether, the combined statistical analysis finds that all three tested factors are significant contributors to scalar diversity. Future work should aim to synthesize all known predictors of scalar diversity reported in the literature (see Section 1), to give us an idea of how much of the total variance in SI rates across scales is now accounted for.

8. Conclusion

In this paper, we replicated scalar diversity on a set of lexical scales drawn from a diverse range of grammatical categories. We then provided experimental evidence for three factors that can capture this observed variation in SI rates. We showed that a production-based measure of how accessible a stronger scalar alternative is can capture scalar diversity. The distinctness of the weak and strong scalar terms, as measured via degree estimates, was also found to be a predictor of SI rates. Lastly, the meaning of the negated stronger scalar term was also shown to play a role.

References

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

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software 67*(1), 1–48.

Beltrama, A. and M. Xiang (2013). Is 'good' better than 'excellent'? An experimental investigation on scalar implicatures and gradable adjectives. In E. Chemla, V. Homer, and G. Winterstein (Eds.), *Proceedings of Sinn und Bedeutung 17*, pp. 81–98.

Davies, M. (2008). The Corpus of Contemporary American English (COCA).

de Marneffe, M.-C. and J. Tonhauser (2019). Inferring meaning from indirect answers to polar questions: The contribution of the rise-fall-rise contour. In M. Zimmermann, K. von Heusinger, and E. Onea (Eds.), *Current Research in the Semantics/Pragmatics Interface*, Volume 36, Questions in Discourse, pp. 132–163. Leiden, The Netherlands: Brill.

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

Drummond, A. (2007). Ibex Farm. `http://spellout.net/ibexfarm`.

Goodman, N. D. and M. C. Frank (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences 20*(11), 818–829.

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

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

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

Horn, L. R. (1989). *A Natural History of Negation*. Chicago: University of Chicago Press.

Kutas, M. and S. A. Hillyard (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature 307*, 161–163.

Lassiter, D. and N. D. Goodman (2017). Adjectival vagueness in a bayesian model of interpretation. *Synthese 194*, 3801–3836.

Levinson, S. C. (2000). *Presumptive Meanings*. MIT Press Ltd.

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

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

Sun, C. and R. Breheny (2021). What the inference task can tell us about the comprehension of scalars and numbers: An investigation of probe question and response bias. Talk presented at Sinn und Bedeutung 26, `https://osf.io/6xmwr`.

Sun, C., Y. Tian, and R. Breheny (2018). A link between local enrichment and scalar diversity. *Frontiers in Psychology 9*, 2092.

Taylor, W. L. (1953). "Cloze procedure": a new tool for measuring readability. *Journalism Quarterly 30*, 415–433.

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

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

Zhang, D. (2021). R-Squared and Related Measures. `https://cran.r-project.org/package=rsq`.