Proceedings of SALT 33: 110–130, 2023

Scalar implicature rates vary within and across adjectival scales*

Helena Aparicio
Cornell University

Eszter Ronai
Northwestern University

Abstract Recent experimental literature has investigated across-scale variation in scalar implicature calculation: lexical scales significantly differ from each other in how likely they are to be strengthened (e.g., *old* → *not ancient* vs. *smart* → *not brilliant*). But in existing studies of this *scalar diversity*, not enough attention has been paid to potential variation introduced by the carrier sentences that scales occur in. In this paper, we carry out the first systematic investigation of the role of sentential context on scalar diversity. Focusing on scales formed by two gradable adjectives, we manipulate the comparison class, specifically whether a noun is likely to have the property described by the scalar adjective (e.g., *brilliant employee* vs. *brilliant scientist*). Our results show within-scale variation: a significant effect of comparison class on the likelihood of scalar implicature calculation. We explain this result in terms of the adjectival threshold distance between the weaker (*smart*) and stronger (*brilliant*) adjective, conditioned on the comparison class (*employee* vs. *scientist*). Our findings also highlight the methodological importance of controlling carrier sentences.

Keywords: experimental pragmatics, scalar implicature, scalar diversity, gradable adjectives, comparison class

1 Introduction: scalar implicature and scalar diversity

In scalar implicature (SI), a weaker statement gets strengthened through hearers' pragmatic reasoning. The utterance in (1a), for example, has the literal lower-bounded meaning in (1b). But if SI is calculated, (1a) gets strengthened to an upper-bounded interpretation, as shown in (1c).

(1) a. The museum is old.
b. The museum is at least old. literal
c. The museum is old, but not ancient. SI

* We thank the reviewers, the audience at SALT 33 and Julian Grove for their invaluable input. This material is partially based upon work supported by the National Science Foundation under Grant No. #BCS-2041312. Authors contributed equally to this work and are listed in alphabetical order.

A standard (neo)-Gricean account of how the strengthened meaning arises is that hearers reason about stronger unsaid alternatives that were also available to the speaker. In the case of the above example, one such alternative utterance is *The museum is ancient* —since this would have been a more informative statement to utter (Quantity Maxim), but the speaker chose not to utter it, its falsity can be inferred (Quality Maxim). Combining the negated stronger alternative with the original utterance results in the SI-strengthened interpretation in (1c) (Grice 1967; Horn 1972).

The same reasoning process is taken to apply to other pairs of lexical items that form a scale, as defined e.g., by asymmetric entailment (Horn 1972). Based on the <*smart*, *brilliant*> scale, for instance, an utterance of (2a) can lead to the strengthened meaning in (2c), which again combines the lower-bounded literal meaning (2b) with the negation of the stronger alternative that was left unsaid (i.e., *The employee is brilliant*).

(2) a. The employee is smart.

b. The employee is at least smart. literal

c. The employee is smart, but not brilliant. SI

As discussed in the following section, previous work has uncovered a great amount of variability in the strength of SI across different types of scales. However, much less is known about the extent to which this observed variability can be traced back to within-scale variation resulting from the carrier sentences used to test different scales (Cf. Degen (2015)). The current study is a first step towards filling this gap.

1.1 Previous work on scalar diversity

An experimental finding that has generated a lot of interest in recent years is that of *scalar diversity*: that different scales differ substantially in how likely they are to lead to SI (van Tiel, Van Miltenburg, Zevakhina & Geurts 2016). For example, the SI-strengthened meaning in (1c) is more likely to arise than the one in (2c). Testing 43 lexical scales, van Tiel et al. (2016) found that the range of SI rates spanned 4% to 100%. The existing body of work investigating scalar diversity has largely concentrated on identifying properties of different lexical scales that can predict how likely they are to lead to SI, thereby explaining the observed across-scale variation (Gotzner, Solt & Benz 2018; Hu, Levy & Schuster 2022; Hu, Levy, Degen & Schuster 2023; Pankratz & van Tiel 2021; Ronai & Xiang 2021, 2022; Sun, Tian & Breheny 2018; van Tiel et al. 2016; Westera & Boleda 2020). However, much less attention has been paid to within-scale variation: namely, how properties of the

sentence a particular scalar term (*old*, *smart*) appears in affect the likelihood of SI calculation, and how these differences relate to across-scale variation, that is, scalar diversity itself.

Different sentential contexts are known to significantly affect the calculation of the more robustly studied *some but not all* SI. An influential investigation comes from Degen (2015), who tested a corpus of 1363 sentences containing the quantifier *some*, and probed whether they are uniformly likely to lead to the calculation of the *some but not all* SI-enriched meaning. Findings showed substantial variation in the robustness of SI calculation, and Degen also identified several properties of the sentential context that predicted SI calculation, such as the partitive structure, determiner strength and discourse accessibility. As mentioned, this study concentrated on the <*some*, *all*> scale; in the domain of scalar diversity, large-scale systematic studies of the role of carrier sentences have not been conducted.

It must be acknowledged that van Tiel et al.'s (2016) landmark paper did test three different sentential contexts for each of the 43 lexical scales in their Experiment 2. For example, SI calculation from the <*old*, *ancient*> scale was tested using the carrier sentences *That {house/mirror/table} is old*. However, van Tiel et al. found no within-scale variation: no pair of sentences for any lexical scale resulted in significantly different rates of SI calculation (p. 148). The three carrier sentences were constructed using the following procedure. A cloze task pre-test was administered with 10 participants, where they were presented with sentences such as *The BLANK is old but it isn't ancient* and had to provide three completions for the blank that would result in a natural-sounding sentence. Of these completions (30 per scale), van Tiel et al. selected three with the goal of ensuring variation, and where possible, picking two high frequency and one low frequency completion.

While a very valuable starting point, van Tiel et al.'s test of carrier sentences was relatively small scale: only 10 participants took part in the pre-test that generated the different sentence frames, and in the main experiment testing SI calculation, each different sentence frame was only seen by 10 participants (for a total of 30 per scale). Subsequent studies on scalar diversity either used van Tiel et al.'s three carrier sentences (Ronai & Xiang 2021), a subset thereof (Sun et al. 2018), or (in the majority of cases) used only a single sentence per scale. It therefore cannot be conclusively ruled out that within-scale variation might still play a role in scalar diversity. In this paper, we conduct a larger scale investigation into the role of sentential context when studying across-scale variation, focusing specifically on the effect of different comparison classes on the likelihood of SI calculation from gradable adjectival scales.

In what follows, we provide a brief introduction to the role of comparison classes in the interpretation of gradable adjectives (Section 1.2). In Section 1.3, we outline the contributions of our paper.

1.2 Adjectival thresholds and comparison classes

In the degree semantics tradition (i.a. Cresswell 1976; von Stechow 1984; Kennedy 1999; Heim 2000; Kennedy & McNally 2005; Kennedy 2007; Syrett, Kennedy & Lidz 2009; Solt & Gotzner 2012), gradable adjectives have been analyzed as relations between an individual x and a degree θ on some abstract adjectival scale associated with the adjective (e.g., intelligence). As seen in (3), the meaning of a gradable adjective states that the degree to which an individual x bears the adjectival property exceeds some adjectival threshold θ, where $\mu_A(x)$ is the measure of x in the scale denoted by the adjective A.

(3) $[\![A]\!] = \lambda\theta_A\lambda x[\mu_A(x) \geq \theta_A]$

The denotation in (3) however does not allow for direct composition of the adjectival predicate with an individual. This compositional problem is fixed by positing a degree morpheme POS (Cresswell 1976; von Stechow 1984; Kennedy 1999; Kennedy & McNally 2005; Kennedy 2007; Grano 2012) that provides a free variable θ_A, whose value is resolved contextually (4a). This silent degree morpheme combines directly with the adjective, saturating the adjective's threshold argument, as seen in (4b):

(4) a. $[\![POS]\!] = \lambda A\lambda x[A(\theta_A)(x)]$
b. $[\![POS\ A]\!] = \lambda x[\mu_A(x) \geq \theta_A]$

The value of the threshold θ_A is thought to be fixed by reasoning about a contextually salient Comparison Class (CC) of individuals that are usually in the extension of the subject NP for predicative adjectives. The value of the θ_A variable is then set such that the CC is partitioned into objects that have the adjectival property, i.e., individuals who have the adjectival property to an equal or a higher degree than θ_A, and those that do not, i.e., those that bear the adjectival property to a lower degree than θ_A. By relativizing the threshold value of the adjective to a CC, it is possible to account for the high degree of context sensitivity displayed by certain gradable adjectives (e.g., the relative adjective *old*), i.e., the fact that an *old cathedral* is significantly older than an *old fruit fly*.[1]

1 Not all gradable adjectives give rise to the same degree of context sensitivity. In particular, absolute adjectives such as *full* are biased towards end-point oriented interpretations. Context-sensitive interpretations of absolute adjectives seem to be limited by how much deviation from the endpoint-oriented interpretation is tolerated in a given (sentential) context. Here we abstract away from the question of whether such context sensitivity should be derived via threshold variability, as is the case for relative adjectives, or by means of other pragmatic mechanisms such as imprecision calculation.

1.3 Overview & contributions of the present study

As reviewed above, experimental studies of across-scale variability in SI calculation, i.e., scalar diversity, have largely implicitly assumed within-scale invariance. While van Tiel et al. (2016) conducted a more limited pre-test comparing three different carrier sentences per scale, they found no differences across them, despite robust findings from i.a., Degen (2015) that sentential context strongly modulates the rate of *some but not all* SI calculation. In this paper, we conduct the first large-scale study of the effect of sentential context on scalar diversity, with an empirical focus on scales formed by gradable adjectives. As we have also discussed above, the interpretation of relative gradable adjectives is dependent on a CC. Therefore, this empirical domain will allow us to investigate the role of sentential context by testing, in particular, the role of different CCs in modulating SI calculation across scales. As our results show, not only is there robust across-scale variation in the adjectival domain (replicating i.a., Pankratz & van Tiel 2021; Gotzner et al. 2018), different CCs also introduce within-scale variation. We consider two hypotheses about the potential role of CCs on SI calculation. Hypothesis 1 links the likelihood of SI calculation to the likelihood of a CC exhibiting the stronger adjectival property. Hypothesis 2 posits that SI rates are instead modulated by the adjectival threshold distance between the weak and strong adjectives, given a CC. We provide a computational model based on Bayesian reasoning that makes explicit the cognitive mechanisms underlying Hypothesis 2.

Our experimental results align with Hypothesis 2 and highlight the methodological importance of controlling carrier sentences. In studies that identified robust scalar diversity effects, but where only one carrier sentence was tested for each scale, some of the observed variation could have, in principle, been driven by sentential context, rather than properties of the lexical scales themselves.

The rest of this paper is structured as follows. In Section 2, we describe two norming experiments we conducted to collect a set of adjectival scales that will form the basis of subsequent experiments, as well as to establish different CCs for each scale. In Section 3, we outline the two hypotheses about the potential role of CCs on SI calculation. The first hypothesis, which links the likelihood of SI calculation to the likelihood of a CC exhibiting the stronger adjectival property, is tested in Section 4. The second hypothesis, which posits that SI rates are instead modulated by the adjectival threshold distance between the weak and strong adjectives, given a CC, is tested in Section 5. Section 6 offers a general discussion of our findings and concludes the paper.

2 Norming studies

In order to test the effect of CCs on SI calculation from adjectival scales, we first needed to collect pairs of adjectives, as well as corresponding CCs. In this section, we report on the two norming studies we conducted in order to do this. Section 2.1 discusses a norming study that tested whether pairs of gradable adjectives pass the relevant semantic tests for scalehood. Section 2.2 discusses the elicitation experiment that gathered two kinds of CCs for each scale: one likely to exhibit the stronger adjectival property, and one unlikely to do so.

2.1 Collecting adjectives

Methods We first gathered adjectival scales that previous work had tested (Gotzner et al. 2018; Pankratz & van Tiel 2021; Ronai & Xiang 2022). From these, we selected ones where the weaker term was a relative gradable adjective;[2] this resulted in a set of 77 scales. As the next step, we normed scales for cancellability and asymmetric entailment (Grice 1967; Horn 1972), by conducting two forced-choice experiments. Experimental tasks were adapted and slightly modified from de Marneffe & Tonhauser (2019). Example (5) illustrates the cancellability test on the <*smart*, *brilliant*> scale: participants saw dialogues such as (5a) or (5b) and had to answer the question "Is Mary's reply to Sue odd?" by clicking "Not odd" or "Odd". Expected answers are given next to the example. Since there is an SI from the weak term (*smart*) to the negation of the strong (*not brilliant*), but this inference is cancellable, the weak-strong order was expected to be judged "Not odd" and the strong-weak order "Odd".

(5) a. Sue: Charlie is smart. — Not odd
Mary: ... and even brilliant!

b. Sue: Charlie is brilliant. — Odd
Mary: ... and even smart!

An example of the test for asymmetric entailment is given in (6), where participants had to answer the question "Does this sentence sound contradictory to you?" with either "Not contradictory" or "Contradictory". Again, expected answers are next to the examples. Since a stronger scalar term (*brilliant*) entails the weaker one (*smart*), but not the other way around, the weak-strong order was expected to be judged "Not contradictory" and the strong-weak order "Contradictory".

2 This included selecting adjectives that Gotzner et al. (2018) had classified as relative. When a classification from a previous scalar diversity study was not available, we adopted the diagnostics of Kennedy & McNally (2005) and Kennedy (2007) to determine whether an adjective is relative vs. absolute.

(6) a. Charlie is smart, but not brilliant. Not contradictory

b. Charlie is brilliant, but not smart. Contradictory

Given that our main interest is the effect of CCs on SI calculation, the norming studies used proper nouns (*Charlie is smart*), or where an inanimate subject was required, pronouns (*It was tasty*). Cancellability and asymmetric entailment were tested between-participants, while the order of scalar terms (weak-strong vs. strong-weak) was manipulated within-participants in each experiment. In addition to the 77 critical items, each experiment contained 2 practice items with feedback about the correct solution, as well as 8 fillers. Fillers were adapted from de Marneffe & Tonhauser (2019) and included sentences that were either clearly "Odd" (*It was expensive... and even cheap!*), "Not odd" (*She is pleasant... and even charming!*), "Contradictory" (*It is open and closed.*), or "Not contradictory" (*Jeff is happy and creative.*).

Participants Native monolingual speakers of American English were recruited on Prolific and compensated $2.50. A total of 80 participants took part, with 40 in each experiment (cancellability and asymmetric entailment). Data from all participants is reported below. The experiments were conducted on the web-based PCIbex platform (Zehr & Schwarz 2018).

Results For a scale to pass the norming, above 60% of the responses needed to be the expected ones for each of the cancellability and the asymmetric entailment test. The 60% threshold was calculated collapsing over the within-participants manipulation. The resulting scale set consists of 48 adjectival scales.[3] It is noteworthy that a relatively high number of scales (29) "failed" the norming tests, despite being used in previous work that had selected items based on researcher intuition and corpus searches. This suggests a need for future studies into the proper criteria for determining scalemate relationships.

3 In our norming studies, we wanted to remain faithful to de Marneffe & Tonhauser's four tested conditions. But while it is clear that (5a) tests for cancellability and (6a)-(6b) together test for asymmetric entailment, it is less obvious what purpose the "[strong]... and even [weak]" (5b) condition serves. Likely relatedly, this condition also produced the lowest rate of the expected response. We tentatively suggest that what may underlie this is that "[strong]... and even [weak]" can be perceived as "Not odd" due to polysemy, specifically if the weaker term is interpreted as having some dimension that is not covered by the stronger term. For instance, *Meg is great... and even nice!* can be interpreted as adding that Meg is kind or entertaining, where *nice* is more than just a weaker scalemate of *great*—leading to a "Not odd" judgment.

2.2 Collecting comparison classes

Methods To gather CCs, we conducted an elicitation experiment. Participants saw stronger scalemates (e.g., *brilliant*, *hilarious*) and were instructed to write down a noun that was likely to have that property. The scalar terms tested were the stronger scalemates from those 48 scales that passed the previous norming study. Since some scales contained the same lexical item as their stronger term (e.g., <*bright*, *brilliant*>, <*smart*, *brilliant*>), the experiment had only 44 items. 2 practice items were included, which provided participants with instructions and sample solutions.

Participants 100 native speaker participants were recruited on Prolific and compensated $2. Data from all participants is reported below. The experiment was run on the web using PCIbex.

Results From the elicited results, we selected two nouns for each scale: one that occurred with high frequency (henceforth "biased") and one that was very infrequent ($\approx$ 1 count; henceforth "neutral"). While selecting CCs, the decision was made to exclude three further scales (<*thin*, *invisible*>, <*pale*, *white*>, <*light*, *white*>) where the elicitation experiment did not provide us with viable candidate nouns for the biased vs. neutral manipulation. Therefore, all subsequent experiments tested 45 adjectival scales.

3 Hypotheses about effect of CC

In this section, we describe two hypotheses about the effect of CCs on SI calculation, which we then test in Sections 4-5. Under Hypothesis 1, the rate of SI calculation is modulated by the likelihood that the CC (*scientist* vs. *employee*) exhibits the stronger scalar property (e.g., brilliance). Under Hypothesis 2, robustness of SI calculation is instead affected by the adjectival threshold distance between the two scalemates (*smart* vs. *brilliant*), given a CC (either *scientist* or *employee*).

3.1 Hypothesis 1: Likelihood

Hypothesis 1 (H1) states that SIs are modulated by the likelihood that the stronger scalemate applies to the noun providing the CC. As seen in Section 1, (neo-)Gricean accounts take SI to arise via listeners' reasoning about what the speaker could have said, but did not (Grice 1967; Horn 1972). For instance, if the speaker chooses to utter that *The employee is smart*, the listener might be compelled to infer that if the speaker did not choose to utter the informationally stronger utterance that *The employee is brilliant* it must be because they do not believe in the truth of that

proposition. Such a chain of reasoning leads to the strengthened interpretation of utterances like *The employee is smart*, where the adjective *smart* is strengthened to *not brilliant*.

Based on this, we can make the following prediction for our CC manipulation. With biased nouns, the stronger adjective was likely to be true of the individual in the CC. For instance, *scientists* are likely to be *brilliant*. The fact that the speaker chose not to utter *brilliant* when describing the *scientist* (but instead used the weaker term *smart*) is then especially meaningful. In other words, since SI arises from the non-utterance of the stronger statement, and with biased CCs, that statement was a priori very likely to be true, we argue that hearers would attribute great importance to the speaker choosing not to say the stronger statement, and they would consequently robustly derive the SI. For neutral CCs, on the other hand, there is higher uncertainty about the applicability of the stronger scalemate: it is less clear that *employees* are *brilliant*. Therefore the listener might be less certain about the reasoning underlying the speaker's utterance choice, which would deter SI calculation. H1 therefore predicts higher rates of SI calculation for biased than for neutral CCs.

3.2 Hypothesis 2: Threshold distance

Hypothesis 2 (H2) claims that SI rates are modulated by the distance between the adjectival threshold of the two scalemates given a CC. Consider Figure 1, which represents the intervals along the adjectival scale denoted by the two scalemates (*smart* in dark purple, *brilliant* in green), as well as the negation of the stronger scalemate (*not brilliant* in orange) and the SI-enriched meaning of the weaker scalemate (*smart and not brilliant*, shown as *smart (SI)* in pale purple). In the display on the left, the overlap between the two scalemates is greater than that observed on the right display, i.e., on the left, there are more degrees that qualify as both *smart* and *brilliant* compared to the right display. This entails that the adjectival thresholds are closer in the former case than in the latter. The closer the thresholds of the two scalemates on the relevant adjectival scale (e.g., smartness), the more overlap between the meanings of the two adjectives, as more individuals can be described with both the weaker *and* the stronger adjective. A direct consequence of the higher overlap between the two scalemates is that the SI-enriched interpretation of the adjective *smart*, i.e., *smart and not brilliant* results in a more strengthened interpretation of the lower scalemate, since there are fewer degrees that fall under both *smart* and *not brilliant* —"smart (SI)" covers a smaller interval on the left than the right in Figure 1.

We argue that this situation discourages SI calculation, as the informational state of the listener (more precisely defined below), rarely warrants such dramatic information gain. One way to conceptualize this is as the listener's counterpart of

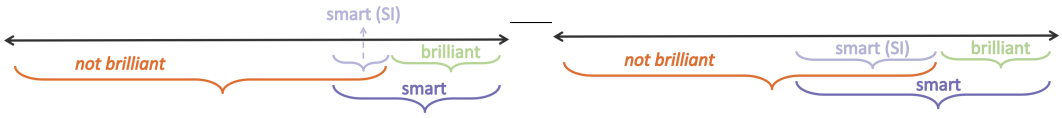

Figure 1 Illustration of potential differences in adjectival threshold distance.

Horn's (1984) R/Q principles: 1) remain faithful to the semantic contribution of the weaker scalemate ("interpret no more than you must", R); 2) strengthen the meaning of the scalemate as much as possible ("interpret as much as you can", Q).

To reflect the fact that listeners' are uncertain about the precise value of adjectival thresholds, we cast these ideas in probabilistic terms, and treat adjectival thresholds as probability distributions ranging over degrees of the relevant scale (Lassiter & Goodman 2013; Qing & Franke 2014). Qualitative predictions are illustrated in Figure 2. The interpretation of the weaker adjective given the negation of the stronger scalemate ($P(smart|\neg brilliant)$), i.e., the SI-enriched interpretation of the weaker adjective, is computed through Bayesian update as shown in (7), where the posterior SI-enriched interpretation of the weaker adjective is proportional to the likelihood, i.e., the probability corresponding to the negated stronger scalemate given the distribution of the weaker adjective, times the prior distribution of the weaker adjective.

(7) $$P(smart|\neg brilliant) \propto P(\neg brilliant|smart)P(smart)$$

Figure 2A illustrates one possible situation where there is high overlap between the strong and weak scalemates. High overlap between $P(smart)$ and $P(brilliant)$, a situation expected to hold when the CC corresponds to a biased noun (e.g., *scientist*), results in an SI-enriched distribution for $P(smart)$ that has very low overlap with its non-SI counterpart (Cf. 2B). Intuitively, high overlap between $P(smart)$ and $P(brilliant)$ entails that, prior to SI calculation, states where *x is smart* is highly

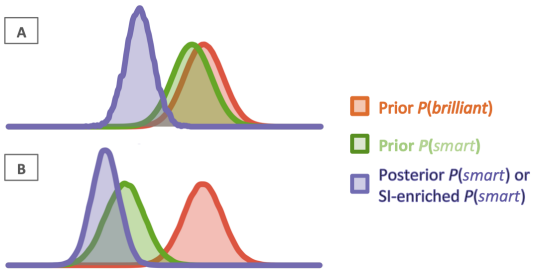

Figure 2 Simulations illustrating qualitative predictions of H2.

probable are also states where *x is not brilliant* is unlikely, where $x \in$ CC. SI calculation therefore has the unwelcome consequence of making initially low probability states highly probable in the posterior. This results in an SI-enriched meaning that's strongly strengthened and distant from the original semantic contribution of the adjective. Given this, high overlap between the prior distributions corresponding to each scalemate should discourage SI calculation. H2 therefore predicts higher SI rates for CCs that lead to larger threshold distances between the thresholds of the relevant scalemates.

4 Evaluating Hypothesis 1: Likelihoods

In order to assess H1, we obtained ratings for how likely the members of the biased and neutral CCs are to bear the adjectival property denoted by the stronger scalemate. We report the results of this experiment in Section 4.1. To determine whether this likelihood is a predictor of SI rates, we used an inference task to test SI calculation; this experiment is reported in Section 4.2. Our results confirm that the biased vs. neutral CCs differ significantly in their likelihood of exhibiting the stronger scalar property. However, contra H1, biased CCs are not more likely than neutral ones to lead to SI.

4.1 Eliciting likelihoods

We experimentally measured the likelihood of the stronger scalar property obtaining with biased vs. neutral nouns. Since the nouns were selected based on an elicitation experiment (Section 2.2) where participants provided nouns likely to have that property, the current experiment served two purposes: 1) to further validate the elicitation results, and 2) to provide us with a continuous, rather than binary measure of likelihood, which we will use to correlate with the likelihood of SI calculation (Section 4.2).

Methods In the experiment, participants were presented with questions such as "On a 0-100 scale, how likely are {employees/scientists} to be brilliant?". Along with this question, they saw a sliding scale with the endpoints labeled "0" and "100" and had to provide their answer by picking a point on that scale. The biased (*scientists*) vs. neutral (*employees*) CC manipulation was tested within-participants. In addition to the 45 critical items, the experiment included 3 practice and 20 filler items. Fillers were constructed to both serve as catch trials and to encourage participants to use the full range of the scale. For instance, we included questions where the expected answer is 0 (*How likely are squares to be round?*), low (*How likely are hamsters to be intelligent?*), or 100 (*How likely are dogs to be mammals?*).

Participants 62 native speakers were recruited on Prolific and compensated approximately $2. One participant was removed due to failure to complete the experimental task; data from 61 is reported below. The experiment was run on PCIbex.

Results & Discussion Results are shown in Figure 3. On average, biased nouns received higher ratings compared to neutral nouns. A linear mixed-effects regression model predicting likelihood rates from the categorical predictor CC BIAS was fitted to the data. The model was maximal, including random intercepts and slopes for participants and items, and the categorical predictor was effect coded. Results show a significant effect of CC BIAS, such that biased nouns were rated significantly higher ($\beta = -24.44$, $SE = 3.45$, $t = -7.0$, $p < 0.001$). The results confirm the validity of the method used for the selection of CCs and provide us with a gradient as opposed to categorical likelihood measure.

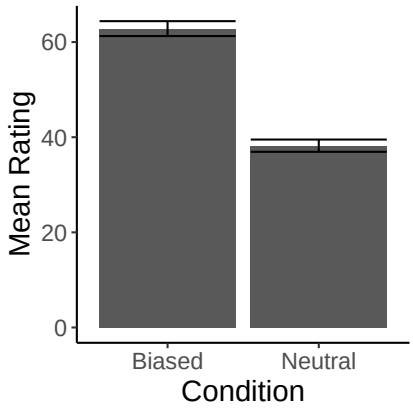

Figure 3 Mean likelihood (and 95% CI) of the biased vs. neutral CC exhibiting the strong adjectival property.

4.2 SI rates

Methods Following van Tiel et al. (2016) (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 employee is smart.*" and were asked the question "Would you conclude from this that Mary thinks the employee is not brilliant?". They responded by clicking "Yes" or "No". A "Yes" answer indicates that the participant has calculated the relevant SI (*smart* → *not brilliant*), while a "No" answer indicates that the participant has not calculated the SI, i.e., they are interpreting *smart* as meaning *at least smart*, compatible with *brilliant*.

The neutral vs. biased CC manipulation was conducted within-participants. For two scales that shared their stronger term (<*bright*, *brilliant*>, <*smart*, *brilliant*> and <*palatable*, *delicious*>, <*tasty*, *delicious*>), we made sure that each participant only saw one of the two relevant scales, i.e., no participant had to make an SI judgment on *not brilliant* or *not delicious* twice. In addition to the 45 critical items, 2 practice and 7 filler items were also included. Fillers contained two antonyms (*wide* $\rightarrow$ *not narrow*, *even* $\rightarrow$ *not odd*). Given that these items had an unambiguously correct answer ("Yes"), they were included to serve as catch trials.

Participants 79 native speaker participants were recruited on Prolific and compensated approximately $2.5. Four participants were excluded for having made four or more mistakes on the filler item catch trials. Four further participants were excluded for taking too long to respond on critical trials, suggesting lack of attention. Data from 71 participants is reported below. The experiment was administered on PCIbex.

Results & Discussion Results are shown in Figure 4. As shown in the plot, neutral CCs gave rise to higher SI rates compared to biased ones. We fitted a logistic mixed-effects regression model to the data, predicting "Yes" vs. "No" responses from the CC BIAS (biased vs. neutral). The model contained random intercepts by items and by participants, as well as by-condition random slopes for both participants and items. The categorical predictor was effect coded. Model outputs confirm a significant effect of CC BIAS, such that Neutral nouns led to significantly more SIs compared to Biased nouns ($\beta = 0.34$, $SE = 0.17$, $z = 2.0$, $p < 0.05$).

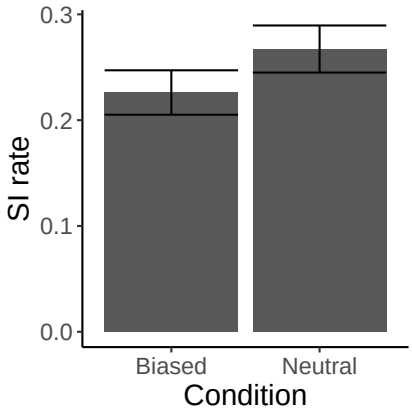

Figure 4 Mean by-CC SI calculation rate (and 95% CI) from the inference task.

This result was also replicated by a by-item analysis (Figure 5), where SI rates were regressed against the likelihood ratings obtained for each item in the likelihood

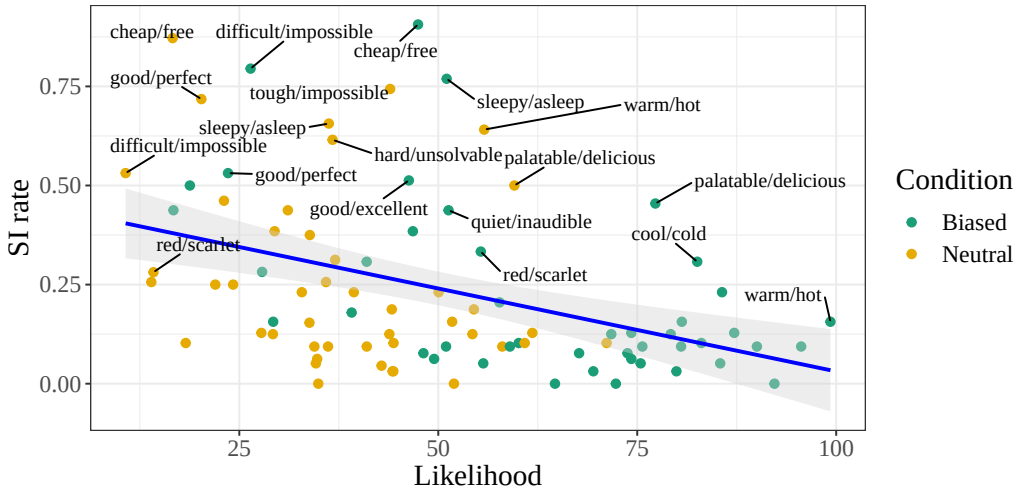

Figure 5 By-item correlation between SI rates and the likelihood of the CC exhibiting the stronger adjectival property. Colors correspond to CCs.

experiment reported in Section 4.1. In line with the logistic mixed-effects model outputs, the likelihood ratings and the SI rates displayed a significant negative correlation ($r = -0.42$, $p < 0.001$), i.e., lower likelihoods yielded higher SI rates. These results are counter to H1, which predicted higher rates of SI calculation with biased CCs than with neutral ones, the opposite of what we found.

5 Evaluating Hypothesis 2: Thresholds

We now proceed to evaluate H2. In Section 5.1 we report results from an experiment that has the goal of obtaining threshold distributions for the same scales and CCs tested in the previous section. We use the elicited threshold data to construct a distance metric, to be more precisely defined in Section 5.1, and examine whether the distance metric is a predictor of the previously obtained SI rates (Cf. Section 4.2). In line with H2, our results show that the SI rates are significantly modulated by the distance between the adjectival thresholds, given a CC.

5.1 Eliciting threshold priors

Methods An experiment was conducted to obtain θ distributions, for both the weaker (8a)-(8b), and stronger adjectives (8c)-(8d), given both neutral and biased CCs. The statement involving the weaker adjective (*smart)* was followed by *possibly*

brilliant in order to block SI calculation. That is, *possibly brilliant* was added to rule out the possibility that participants would calculate the SI from *The employee is smart* and provide threshold estimates given an enriched *The employee is smart, but not brilliant* meaning.

(8) a. The employee is smart, possibly brilliant. neutral, weak
b. The scientist is smart, possibly brilliant. biased, weak
c. The employee is brilliant. neutral, strong
d. The scientist is brilliant. biased, strong

Participants were presented with an utterance such as one of the ones from (8), along with a sliding scale with endpoints labeled "0" and "100". On the same screen as the utterance and the sliding scale, participants were asked the question "On a 0-100 scale, how smart is the {scientist/employee}?". They provided their judgement by picking a point on the sliding scale. The task questions ("On a 0-100...") always relied on the weaker term from the scale. The weak vs. strong and neutral vs. biased manipulations were conducted between-participants. In addition to the 45 critical items, the experiment included 3 practice and 5 fillers items. The latter included antonyms (e.g., *The table is clean. On a 0-100 scale, how dirty is the table?*) and served as catch trials.

Participants 240 native speakers were recruited on Prolific; 60 people participated in each between-participants condition. Four participants were removed because they failed to use the lower half of the response scale; data from 236 participants is reported below. The experiment was run on PCIbex.

Results & Discussion On average, the mean ratings were higher for strong scalemates compared to weak scalemates for both neutral and biased CCs. This was confirmed by a series of t-tests comparing threshold ratings for the weak and the strong scalemate within CC (neutral: $t(44) = 3.57$, $p < 0.001$; biased: $t(44) = 5.1$, $p < 0.001$).

To determine the effect of threshold distance on SI rates, we computed a distance score D with the goal of quantifying the distance between the elicited thresholds of two scalemates on a shared adjectival scale. We first computed the metric d_n, which captures the distance between the two scalemates' thresholds within a CC. As seen in equation (9), d_n was computed by subtracting the mean μ of the threshold ratings for the weak scalemate w_n from the strong scalemate s_n, where n stands for a particular CC. The difference between the means was subsequently standardized by dividing by the product of the standard deviations (σ) of the relevant random variables (see equation (9); Cf. Toscano & McMurray 2010 for a similar distance metric).

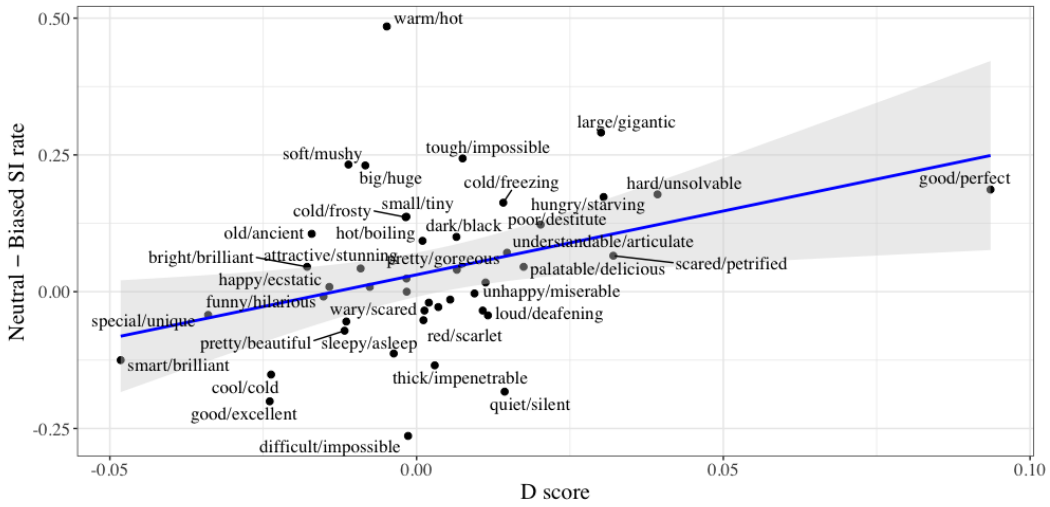

Figure 6 By-item correlation between the neutral vs. biased difference in SI rates and the *D* scores, which index adjectival threshold distance.

(9) $$d_n = (\mu_{s_n} - \mu_{w_n})/\sigma_{s_n}\sigma_{w_n}$$

In order to determine potential CC effects within an adjectival scale, a second difference score *D* was obtained by subtracting the difference score for biased nouns from the difference score obtained for neutral nouns, see equation (10).

(10) $$D = d_{n_{neut.}} - d_{n_{bias.}}$$

Figure 6 plots the difference in SI rates obtained in the neutral vs. biased conditions in the inference task reported in Section 4.2, against the *D*-score. Visual inspection of the plot suggests that there exists a positive relationship between threshold distance and SI rates, a trend that was statistically confirmed ($r = 0.36$, $p < 0.02$), such that scales for which the SI difference between the neutral and the biased conditions was larger tended to have a higher *D*-score. The current results therefore constitute evidence for H2: the distance metric, which is itself a function of the CC manipulation used in the studies reported in this paper, predicts higher SI rates for scales whose scalemates are perceived to have more distant thresholds.

6 General Discussion & Conclusion

In this paper we revisited the question of whether the robustness of SI calculation shows not only across-scale, but also within-scale variation. Testing 45 different

lexical scales, we indeed found both that they differ from each other in how likely they are to give rise to SI (replicating the scalar diversity phenomenon) and that different sentential contexts modulate this likelihood. Specifically, we focused on lexical scales formed by gradable adjectives and manipulated what CC the adjectives were to be interpreted relative to. For each scale, two CCs were established: a biased CC, where the noun was likely to have the adjectival property described by the stronger scalemate (e.g., *scientist* for the <*smart*, *brilliant*> scale) and a neutral CC, where this likelihood was not especially high (e.g., *employee* for <*smart*, *brilliant*>). We found a significant effect of the biased vs. neutral CC manipulation across the board: neutral CCs led to higher rates of SI calculation.

Our finding that sentential contexts introduce within-scale variation in SI rates is expected given previous work such as Degen's (2015) on the <*some*, *all*> scale. But it seemingly goes against van Tiel et al.'s (2016) original scalar diversity study, which found no difference across carrier sentences within the same scale. At the same time, a number of key differences between our work and van Tiel et al.'s may be able to explain this discrepancy. First, our experiments recruited a larger number of participants both for establishing the different carrier sentences and for testing their effect on SI calculation (see Section 1.1). Second and more crucially, we specifically focused on making one carrier sentence per scale "biased", while van Tiel et al.'s method of eliciting these sentences merely asked participants for a natural-sounding completion, which likely gave more neutral results overall. One potential explanation is therefore that in the context of robust scalar diversity, the effect of different sentential contexts is quite nuanced and hard to identify experimentally.

We discussed two hypotheses about what might underlie the effect of CCs on within-scale SI rate variation. According to H1, SI calculation is directly affected by how likely the CC is to exhibit the stronger scalar property. Since biased CCs are more likely to do so, and SI arises from the non-utterance of a stronger alternative, we argued that hearers would attribute great importance to the speaker's choice to use the less informative weaker scalemate, leading to higher rates of SI. However, the experimental results found the opposite effect. Under H2, on the other hand, SI rates are modulated by the distance between the adjectival threshold of the weaker vs. stronger scalemate, given a CC.[4] Our findings were in line with this hypothesis.

We must note that even though H1 was disconfirmed by the data, likelihood still had a direct effect, except in the opposite direction to what we had predicted: biased CCs led to less, rather than more SI. We want to suggest that this effect of likelihood can in fact be reduced to threshold distance and accommodated under H2. We argue that likelihood perception is, among other things, a by-product of the fact that, prior

4 Though we have spelled out H2 in probabilistic terms (Section 3.2), it is also compatible with the non-probabilistic semantic distance proposal of van Tiel et al. (2016) (going back to Horn 1972), who found higher SI rates *across* scales with more semantically distant scalemates.

to (potential) SI calculation, the weaker scalemate is more likely to apply than the stronger scalemate (due to their asymmetric entailment relation). Therefore, the larger the distance between the two scalemates, the less likely the stronger scalemate is relative to the weaker scalemate, as more degrees and potentially individuals in the CC can only be described with the weaker scalemate. As discussed in Section 3.2 —see Figures 1-2 —larger distance leads to higher SI rates. This also means that the likelihood of the weak scalemate should interact with that of the strong scalemate and CC, such that the likelihood of the weak scalemate should be significantly higher than that of the strong scalemate for neutral CCs compared to biased ones.

Lastly, let us touch on two methodological conclusions that emerge from our study. First, in the norming experiment for establishing whether two adjectives form a scale, over a third of the tested items ended up being excluded. This is despite the fact that all of them had been used in previous studies, which had selected scales based on prior literature, researcher intuition and corpus searches. While questions remain about how to experimentally implement the relevant semantic tests for scalehood (see also fn. 3 in Section 2.1) and what cutoff to employ for counting a scale as having "failed" those tests, it is nevertheless informative that so many (purported) scales needed to be excluded, suggesting the need for future research.

Second, the question arises what implication our main finding —that sentential context significantly affects likelihood of SI for a large number of different scales —has for previous studies of scalar diversity. In principle, it is possible that the uncontrolled effect of CCs had introduced a confound in prior work. But for this to necessitate the reinterpretation of previous findings, there would need to be a systematic bias in previous experiments, such that some scales had been tested with what would count as a neutral CC, and some others with what would count as a biased CC. Hypothetically, if a prior study had observed that Scale 1 is less likely to lead to SI than Scale 2, but Scale 1 was tested with a biased and Scale 2 with a neutral noun, then the difference in SI rates could have arisen as an artifact of the CCs, and not due to properties of the lexical scales themselves. However, this situation is not especially likely to have been the case pervasively enough to explain all of scalar diversity. The fact that prior work has successfully identified properties of lexical scales as predictors of scalar diversity also suggests that the across-scale variation in SI rates cannot be reduced to an illusion arising from uncontrolled carrier sentences. At the same time, future work should pay closer attention to controlling carrier sentences, and testing a larger variety of them, to help us gain a fuller understanding of how much variation can be attributed to the identity of lexical scales vs. contextual cues —for similar arguments, see also Degen (2021). Finally, our paper proposes a mathematical formalization for the effect of semantic distance on SI calculation. As a future direction, the current proposal can therefore be extended to make finer grained by-item predictions about SI likelihood for a given scale and CC.

References

Cresswell, Max J. 1976. The semantics of degree. In Barbara Partee (ed.), *Montague grammar*, 261–292.

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. doi:10.3765/sp.8.11.

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.

Gotzner, Nicole, Stephanie Solt & Anton Benz. 2018. Scalar diversity, negative strengthening, and adjectival semantics. *Frontiers in Psychology* 9. 1659.

Grano, Thomas Angelo. 2012. *Control and restructuring at the syntax-semantics interface*. The University of Chicago.

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

Heim, Irene. 2000. Degree operators and scope. In *Semantics and Linguistic Theory (SALT) 10*, 40–64.

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

Hu, Jennifer, Roger Levy, Judith Degen & Sebastian Schuster. 2023. Expectations over Unspoken Alternatives Predict Pragmatic Inferences. *Association for Computational Linguistics* 11. 885–901. doi:10.1162/tacl_a_00579.

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

Kennedy, Christopher. 1999. *Projecting the adjective: The syntax and semantics of gradability and comparison*. Garland.

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

Kennedy, Christopher & Louise McNally. 2005. Scale structure, degree modification, and the semantics of gradable predicates. *Language* 81(2). 345–381. doi:10.1353/lan.2005.0071.

Lassiter, Daniel & Noah D Goodman. 2013. Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Semantics and Linguistic Theory (SALT 23)*, 587–610.

de Marneffe, Marie-Catherine & Judith Tonhauser. 2019. Inferring meaning from indirect answers to polar questions: The contribution of the rise-fall-rise contour.

In Malte Zimmermann, Klaus von Heusinger & Edgar Onea (eds.), *Current research in the semantics/pragmatics interface*, vol. 36, Questions in Discourse, 132–163. Leiden, The Netherlands: Brill. doi:10.1163/9789004378322_006.

Pankratz, Elizabeth & Bob van Tiel. 2021. The role of relevance for scalar diversity: a usage-based approach. *Language and Cognition* 13(4). 562–594. doi:10.1017/langcog.2021.13.

Qing, Ciyang & Michael Franke. 2014. Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. In *Semantics and Linguistic Theory (SALT 24)*, 23–41.

Ronai, Eszter & Ming Xiang. 2021. Exploring the connection between Question Under Discussion and scalar diversity. *Linguistic Society of America (LSA)* 6(1). 649–662. doi:10.3765/plsa.v6i1.5001.

Ronai, Eszter & Ming Xiang. 2022. Three factors in explaining scalar diversity. *Sinn und Bedeutung* 26. 716–733.

Solt, Stephanie & Nicole Gotzner. 2012. Who here is tall? Comparison classes, standards and scales. In *International Conference on Linguistic Evidence*, 79–83.

von Stechow, Arnim. 1984. Comparing semantic theories of comparison. *Journal of Semantics* 3(1-2). 1–77. doi:10.1093/jos/3.1-2.1.

Sun, Chao, Ye Tian & Richard Breheny. 2018. A link between local enrichment and scalar diversity. *Frontiers in Psychology* 9.

Syrett, Kristen, Christopher Kennedy & Jeffrey Lidz. 2009. Meaning and Context in Children's Understanding of Gradable Adjectives. *Journal of Semantics* 27(1). 1–35. doi:10.1093/jos/ffp011.

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

Toscano, Joseph C. & Bob McMurray. 2010. Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science* 34(3). 434–464. doi:10.1111/j.1551-6709.2009.01077.x.

Westera, Matthijs & Gemma Boleda. 2020. A closer look at scalar diversity using contextualized semantic similarity. *Sinn und Bedeutung* 24(2). 439–454. doi:10.18148/sub/2020.v24i2.908.

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

Helena Aparicio
203 Morrill Hall
159 Central Ave
Ithaca, NY 14850
haparicio@cornell.edu

Eszter Ronai
2016 Sheridan Rd
Room 205
Evanston, IL 60208
ronai@northwestern.edu