*Journal of Semantics*, 00, 2025, 1–30
https://doi.org/10.1093/jos/ffaf002
**Article**

OXFORD

# Scalar implicature rates vary within and across adjectival scales

**Helena Aparicio**[1,‡] **and Eszter Ronai**[2,‡,*]

[1]Department of Linguistics, Cornell University, 203 Morrill Hall, 159 Central Ave, Ithaca, New York, NY 14853, United States
[2]Department of Linguistics, Northwestern University, 2016 Sheridan Rd, 60208, Evanston, IL, United States
*Corresponding author: Department of Linguistics, Northwestern University, 2016 Sheridan Rd, 60208, Evanston, IL, USA. E-mail: ronai@northwestern.edu

‡Helena Aparicio and Eszter Ronai share first authorship and are listed in alphabetical order.

## Abstract

Recent experimental literature has investigated across-scale variation in scalar implicature calculation, probing why lexical scales differ from each other in their likelihood of being strengthened (e.g. *old* → *not ancient* v. *smart* → *not brilliant*). But in existing studies of this *scalar diversity*, less attention has been paid to potential variation introduced by the carrier sentences that scales occur in. In this paper, we carry out a systematic investigation of the role of sentential context on scalar diversity, focusing on scales formed by two gradable adjectives. We find within-scale variation: different subject nouns (e.g. *The employee is smart* v. *The scientist is smart*) have a significant effect on how robustly a scalar implicature arises. We then explore the relationship between a noun's prior likelihood of exhibiting the stronger adjectival property (e.g. brilliance) and the rate of implicature calculation, and find that they are negatively correlated. We also test whether a previously identified factor in scalar diversity, adjectival threshold distance between the weaker (*smart*) and stronger (*brilliant*) adjective, is sensitive to the subject noun manipulation, but do not find evidence for this. In addition to their theoretical import, our findings also highlight the methodological importance of controlling carrier sentences.

*Keywords:* scalar implicature; scalar diversity; gradable adjectives; comparison class; likelihood priors

## 1 Introduction: scalar implicature

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

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

**Received** 5 January 2024; **revised** 28 January 2025; **accepted** 28 January 2025

A standard (neo)-Gricean account of how the strengthened meaning arises is that hearers reason about informationally stronger unsaid alternatives that were also available to the speaker. In the case of the above example, a relevant alternative utterance to (1-a) is *The museum is ancient*. This alternative can be taken to be more informative than *The museum is old* because it asymmetrically entails it (Horn 1972). Since *The museum is ancient* would have been a stronger statement to utter, the speaker should have uttered it in place of (1-a) if it were true (and the speaker knew so). Therefore, because they did not utter the alternative, its falsity (*The museum is not ancient*) can be inferred. This process can be viewed as the interaction of the Gricean submaxims Quantity-1 ("Make your contribution as informative as is required (for the current purposes of the exchange)") and Quality-1 ("Do not say what you believe is false") (Grice 1967). Combining the negated stronger alternative (*The museum is not ancient*) with the original utterance (*The museum is old*) results in the SI-strengthened interpretation in (1-c).

The above example of SI is based on *old* and *ancient* forming a lexical scale, which, as mentioned, can be defined via asymmetric entailment. The same pragmatic reasoning process illustrated on <*old*, *ancient*> can also apply to other scales. Based on <*smart*, *brilliant*>, for instance, an utterance of (2-a) can lead to the strengthened meaning in (2-c), which again combines the lower-bounded literal meaning (2-b) 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-strengthened

As discussed in the following section (Section 1.1), previous work has uncovered a great amount of variability in the strength of SI across different types of scales (i.a. Baker *et al.* 2009; Beltrama and Xiang 2013; Doran *et al.* 2012; van Tiel *et al.* 2016). However, relatively little is known about how this observed variability is modulated by different sentential contexts, which are known to result in within-scale variation in SI calculation (Degen 2015). The current study is a first step towards filling this gap.

## 1.1 Previous work on across- and within-scale variation

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. For example, across studies, the SI-strengthened meaning in (1-c) is more likely to arise than the one in (2-c) (Gotzner *et al.* 2018b; Pankratz and van Tiel 2021). Such effects also go far beyond just these two particular scales; in the first comprehensive study of scalar diversity, van Tiel *et al.* (2016) tested 43 lexical scales and found that the range of SI rates spanned 4% to 100%. This effect is unexpected under prior assumptions that findings about a single scale should generalize to all cases of SI (see the *uniformity assumption*, van Tiel *et al.* 2016, p. 139). As such, a growing number of experimental studies has tried to explain why scalar diversity arises. The approach taken in most existing studies is to try to identify relevant properties that lexical scales differ in, which can predict how likely they are to lead to SI, ultimately explaining the observed across-scale variation. Such properties include how distinct the weak and the strong scalar terms are on a scale (van Tiel *et al.* 2016; Westera and Boleda 2020), how certain a hearer can be about the identity of the stronger alternative given a weaker scalar (Ronai and Xiang 2020)—though this effect is possibly not driven by the particular stronger lexical items, but rather by a so-called meaning-based notion of alternatives (Hu *et al.* 2023, 2022)—the polarity or extremeness of adjectival scales (Beltrama and Xiang 2013; van Tiel *et al.* 2016), a scale's propensity for undergoing other semantic or pragmatic enrichment (Gotzner *et al.* 2018b; Sun *et al.* 2018), a scale's

relation to features of the discourse context (Ronai and Xiang 2021), or the relevance of the SI itself (Pankratz and van Tiel 2021). However, while trying to explain differences in SI calculation across different scales, this existing body of work has paid less attention to within-scale variation: namely, how properties of the sentence a particular weaker scalar term appears in affect the likelihood of SI calculation, and how this might relate to scalar diversity.

Different sentential contexts are known to significantly affect the calculation of the more robustly studied *some but not all* SI.[1] An influential investigation of <*some*, *all*> 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 (2015) also identified several properties of the sentential context that predicted SI calculation, such as the partitive structure, determiner strength and discourse accessibility. This empirical finding was later replicated by Hu *et al.* (2023) and extended to <*or*, *and*> by Li *et al.* (2021). Sun *et al.* (2023) have recently taken a similar approach to Degen (2015) and tested SI calculation from 28 different lexical scales in a corpus of Twitter data. They found that the scalar diversity effect was reduced compared to studies that had used a limited number of manually constructed sentential contexts (Sun *et al.* 2018; van Tiel *et al.* 2016). At the same time, the observed across-scale variation in SI rates was explained by the same scale-intrinsic factors that had been correlated with scalar diversity in prior work, and these factors in fact explained a similar amount of variance. Sun *et al.* (2023)'s study did not investigate what properties of the sentential context make SI calculation more or less likely, either on average or for each different lexical scale. The present study aims to more directly address this question, by systematically manipulating carrier sentences in order to further explore within-scale variation in the presence of across-scale variation. Specifically, we focus on the effect of different subject nouns on the likelihood of SI calculation from gradable adjectival scales.

A direct manipulation of carrier sentences was done by van Tiel *et al.* (2016), who tested 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.* (2016) 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), the authors 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.* (2016)'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). These aspects of the study might explain the null result. Since van Tiel *et al.* (2016), subsequent studies on scalar diversity either used their three carrier sentences (Ronai and Xiang 2020), a subset thereof (Sun *et al.* 2018), or (in the majority of cases) used only a single sentence per scale—with, as mentioned, the notable exception of Sun *et al.* (2023). This leaves open the possibility that there is a systematic way in which sentential context interacts with scalar diversity, which is what the current paper aims to probe.

[1] The robustness of SI calculation from *some* is also known to be impacted by features of the experiment (Geurts and Pouscoulous 2009; Jasbi *et al.* 2019, i.a.), the discourse context (Degen 2013; Ronai and Xiang 2020; Zondervan *et al.* 2008, i.a.), or participant characteristics (Fairchild and Papafragou 2021, i.a.), among other factors, but these are not the focus of the current paper.

To set the stage for our own experimental manipulation, we now turn to a brief introduction to gradable adjectives, and the role of comparison classes in their interpretation (Section 1.2). Section 1.3 outlines the contributions of our paper in more detail.

## 1.2 Adjectival thresholds

In the degree semantics tradition (i.a. Cresswell 1976; Heim 2000; Kennedy 2007; Kennedy and McNally 2005; Kennedy 1999; Solt and Gotzner 2012a,b; Syrett *et al.* 2009; Stechow 1984), gradable adjectives have been analysed as relations between an individual $x$ and a degree $\theta$ 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 equals or exceeds some adjectival threshold $\theta$, 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 that provides a free variable $\theta_A$, whose value is resolved contextually (4-a). This silent degree morpheme combines directly with the adjective, saturating the adjective's threshold argument, as seen in (4-b):

(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 $\theta_A$ is thought to be fixed by reasoning about a contextually salient comparison class (CC) of individuals that often corresponds to the extension of the subject NP for predicative adjectives. The value of the $\theta_A$ 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 $\theta_A$, and those that do not, i.e. those that bear the adjectival property to a lower degree than $\theta_A$. 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. *old*), i.e. the fact that an *old cathedral* is significantly older than an *old fruit fly*.[2]

In order to determine the CC, comprehenders make use of different types of linguistic and extra-linguistic information. Previous work has found that both children and adults make use of the linguistic context, e.g. the syntactic position in which the adjective appears (Tessler *et al.* 2020), the identity of the NP that the adjective takes as a semantic argument (Barner and Snedeker 2008; Ebeling and Gelman 1994), the adjective's polarity (Tessler and Goodman 2022), the makeup of the visual context (Barner and Snedeker 2008; Foppolo and Panzeri 2013; Gotowski and Syrett 2020; Syrett *et al.* 2010), general world knowledge (Ebeling and Gelman 1988; Gelman and Ebeling 1989; Tessler and Goodman 2022), or the observed distribution of degrees in the CC (Cremers 2022; Lassiter and Goodman 2013), among other cues. In the experiments reported below, we manipulate the subject noun (e.g. *The{employee, scientist}is brilliant*) to induce changes to the CC that will ultimately result in different values for the threshold used for the interpretation of the adjective. We detect such changes by eliciting participants' judgments about the degree to which the individual

[2] Not all gradable adjectives give rise to the same degree of context sensitivity. In particular, absolute adjectives such as *full* are biased towards endpoint-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 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.

denoted by the subject noun bears the adjectival property, the assumption being that higher thresholds should elicit higher degrees on average.

### 1.3 Overview and contributions of the present study

As reviewed above, experimental studies of across-scale variability in SI calculation, i.e. scalar diversity, have largely set aside potential within-scale variation. Though 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 the *some but not all* SI calculation. And while Sun *et al.* (2023) found that across-scale variation is attenuated (though still present) when multiple different corpus occurrences of a scalar term are tested, this study did not focus on what specific aspects of carrier sentences impact the likelihood of SI calculation, and how this relates to scalar diversity itself. In this paper, we systematically manipulate carrier sentences to test their effect 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 context-dependent. Therefore, this empirical domain will allow us to investigate the role of sentential context by testing, in particular, how different subject nouns modulate SI rates. As our results show, not only is there robust across-scale variation in the adjectival domain (replicating i.a., Gotzner *et al.* 2018b; Pankratz and van Tiel 2021), different subjects that the adjectives are predicated of also introduce within-scale variation.

Our primary hypothesis about the potential role of sentential context on SI calculation links the likelihood of SI calculation to the likelihood of a noun exhibiting the stronger adjectival property (Hypothesis 1). This could manifest as a negative relationship: if a stronger statement (e.g. *The ruins are ancient*) was a priori likely to be true, hearers might be disinclined to calculate an SI that would contradict it (*The ruins are old but not ancient*) (Degen *et al.* 2015; Tsvilodub *et al.* 2023). At the same time, we argue that a positive relationship is also possible: the speakers' non-utterance of an a priori likely statement might be especially salient to the hearer, encouraging SI calculation. Our experimental results suggest that sentential context indeed affects SI calculation rates; specifically, we find a negative correlation between how likely a noun is to have the adjectival property and the likelihood of SI. Motivated by this empirical finding, in Hypothesis 2 we explore whether a known correlate of SI rates, semantic distance (Horn 1972; van Tiel *et al.* 2016), also varies within-scales, depending on the subject noun. Our results do not reveal evidence for this, and instead we find that semantic distance continues to be a predictor of scalar diversity irrespective of the subject noun.

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 subject nouns for each scale. In Section 3, we outline and test our main hypothesis (Hypothesis 1) about the potential role of sentential context in SI calculation, namely that SI calculation is modulated by prior likelihood. In Section 4, we test whether sentential context has an impact on semantic distance and its role in predicting SI calculation (Hypothesis 2). Section 5 offers a discussion of our findings, including their methodological consequences. Section 6 concludes.

## 2 Norming studies

In order to test the effect of sentential context on SI calculation from adjectival scales, we first needed to collect pairs of adjectives, as well as corresponding potential subject nouns. 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 subjects for each scale: one likely to exhibit the stronger adjectival property, and one unlikely to do so.

## 2.1 Experiment 1: Collecting adjectives

### 2.1.1 Methods

**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 and Schwarz 2018).

**Materials and Procedures** We first gathered adjectival scales that previous work had tested (Gotzner *et al.* 2018b; Pankratz and van Tiel 2021; Ronai and Xiang 2022). From these, we selected ones where the weaker term was a relative gradable adjective[3]; this resulted in a set of 77 scales. As the next step, we normed these 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 and Tonhauser (2019). Example (5) illustrates the cancellability test on the <*smart*, *brilliant*> scale: participants saw dialogues such as (5-a) or (5-b) 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".

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

Given that our main interest in this paper is the effect of sentential context on SI calculation, the norming studies used neutral contexts: subjects were 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 v. 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 and 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*).

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

### 2.1.2 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 tests.[4] In calculating whether a scale reached the 60% threshold, trials from the within-participants conditions were analysed together, that is, (5-a) was analysed together with (5-b) and (6-a) together with (6-b). 48 adjectival scales passed the norming.[5] It is noteworthy that a relatively high number (29) of scales did not, despite being used in previous work that had selected items based on researcher intuition and corpus searches.

One reason for this might be the presence of lexical semantic factors not having to do with scalehood *per se*. While we wanted to remain faithful to de Marneffe and Tonhauser (2019)'s four tested conditions, and it is clear that (5-a) tests for cancellability and (6-a)–(6-b) together test for asymmetric entailment, it is less obvious what purpose the "[strong]... and even [weak]" (5-b) condition serves. Likely relatedly, this condition also produced the lowest rate of the expected response. We tentatively suggest that what may underlie this result 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.[6] Interestingly, one prior study that has directly looked at polysemy in the context of scalar diversity, Sun *et al.* (2018), has used a task near-identical to our condition (6-b) to operationalize polysemy. Yet in our own data, this condition did not seem to exclude disproportionately many scales: the expected answer was given 75% of the time for (5-a), 49% for (5-b), 88% for (6-a), and 73% for (6-b). Overall, our results suggest a need for future studies into the proper criteria for determining scalemate relationships, as well as into different ways of identifying polysemy, and isolating the two from one another.

## 2.2 Experiment 2: Collecting subject nouns

### 2.2.1 Methods

**Participants** 100 native monolingual speakers of American English were recruited on Prolific and compensated $2. Data from all participants is reported below. The experiment was run on the web using PCIbex.

**Materials and Procedures** To gather subject nouns, 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 adjectives from the 48 scales that remained after 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. A total of 2 practice items were included, which provided participants with instructions and sample solutions.

### 2.2.2 Results

From the elicited results, we selected two nouns for each scale: one that occurred with high frequency and one that was very infrequent (≈ 1 count). In what follows, we refer to the high frequency noun as the "biased" subject, since these are the nouns robustly provided

[4] We selected the 60% cutoff as a way to make our analysis similar to de Marneffe and Tonhauser (2019), who specified their criterion as the "majority of judgments" being the expected one. The cutoff was determined before data collection.

[5] For the list of adjectives that did not pass the norming, see Appendix I.

[6] The reader may wonder whether it is a shortcoming of the norming that it potentially excludes not just purported scales that do not actually pass the test(s) for scalehood, but also those with a polysemous scalar term. Given that polysemy is a not-yet fully understood factor in scalar diversity (Sun *et al.* 2018), and our goal in this study is to test the role of other predictors, we see it as desirable to restrict our attention to non-polysemous scales by employing a more conservative norming criterion.

as likely to have the adjectival property in question. The low frequency noun, in turn, will be referred to as the "neutral" subject. Since they showed up in the elicitation, these nouns are compatible with the adjective, but they are not especially likely to exhibit the relevant property. To give one example, for the <*smart*, *brilliant*> scale, *scientist* was selected as the biased subject, and *employee* as the neutral subject.

While selecting nouns, 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 v. neutral manipulation. Therefore, all subsequent experiments tested 45 adjectival scales.

## 3 Hypothesis 1: Likelihoods

Hypothesis 1 (H1) states that SIs are modulated by the likelihood that the stronger scalemate applies to an individual in the extension of the subject noun. In particular, existing literature (Degen *et al.* 2015; Tsvilodub *et al.* 2023) on the role of prior likelihood in SI calculation has put forward the hypothesis that we summarize in (7).

(7) **Hypothesis 1a (H1a)**: The more likely a stronger alternative statement is to hold, the less likely hearers are to calculate an SI to the negation of that stronger alternative.

As i.a. Degen *et al.* (2015) note, H1a is a prediction of the Rational Speech Acts framework (RSA, Frank and Goodman 2012). Within RSA, pragmatic inferences are modeled via Bayesian reasoning, with prior beliefs therefore playing an important role in interpretation. Degen *et al.* (2015) tested H1a on the <*some*, *all*> scale, comparing utterances like *Some of the marbles sank* with *Some of the feathers sank*. Based on world knowledge, hearers know that the stronger alternative (*All of the X sank*) is more likely to be true in the case of marbles than feathers; the prediction under H1a, then, is that the *Some but not all X sank* SI will not—or is at least less likely to—arise in the marble case. Degen *et al.* (2015) collected experimental judgments on the prior likelihood that a certain number (including all) of the X (marbles, feathers, etc.) sank, and correlated these with SI judgments from the corresponding utterances containing *some*. They found a significant effect showing that the more likely the stronger alternative was to be true a priori, the less likely the SI calculation.[7] Notably, however, the effect of prior likelihood found by Degen *et al.* (2015) was much smaller than that predicted by the standard RSA, prompting the authors to propose an extension of the model; by studying adjectives instead of a quantificational scale, which are likely to display more context sensitivity, we may be able to see more robust effects of likelihood.

Following up on Degen *et al.* (2015), Tsvilodub *et al.* (2023) tested the role of prior likelihood in the calculation of SI from the <*some*, *all*> and <*or*, *and*> scales, but only found an effect for <*some*, *all*>. The authors take this as evidence that the *not both* inference from *or* is not an SI, though they also note that by-participant variability could underlie the null result. Assuming that findings about the role of prior likelihood in the calculation of the *some but not all* SI extend to SI calculation from a larger variety of (adjectival) scales, we can make the following prediction for our subject manipulation under H1a. Since in the biased subject condition, nouns are likely to exhibit the stronger adjectival property, e.g. a

[7] This is in contrast to Geurts (2010), who argued that implausibility is not sufficient to stop an SI from going through, based on his judgment that a sentence like (8) still conveys that not all of Cleo's marbles sank.

(8) Cleo threw all her marbles in the swimming pool. Some of them sank to the bottom. (ex. 60 from Geurts 2010, p. 157).

*scientist* is likely to be *brilliant*, hearers will be less inclined to derive the SI that would be counter to this, namely that *The scientist is smart but not brilliant*. Therefore, the biased subject condition should show a reduction in SI rates.

At the same time, even though prior work has posited a negative relationship between likelihood and SI (or no relationship, see fn. 7), the opposite prediction can also be made. This alternative hypothesis is stated in (9).

(9) **Hypothesis 1b (H1b)**: The more likely a stronger alternative statement is to hold, the more likely hearers are to calculate an SI to the negation of that stronger alternative.

The reasoning in favor of H1b is as follows. As we saw 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). The experimental manipulation might have the following effect. With biased subjects, the stronger adjective is likely to be true of an individual in the extension of the noun, e.g. *brilliant scientist*. The fact that the speaker chose not to utter *brilliant* when describing the *scientist* (but instead used the weaker term *smart*) is especially meaningful. That is, since SI arises from the non-utterance of the stronger statement, hearers may attribute great importance to the speaker choosing not to say the stronger, a priori likely statement; consequently, they should robustly derive the SI. For neutral subjects, on the other hand, there is a priori 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. H1b therefore predicts higher rates of SI calculation for biased than for neutral subjects.

In Section 3, we test the role of prior likelihood in SI calculation on 45 different lexical scales. In doing so, we contrast H1a, which predicts a negative relationship between the prior likelihood of the stronger alternative statement and the robustness of SI calculation from the weaker one, and H1b, which predicts a positive relationship. In order to assess H1, we obtained ratings for how likely the biased and neutral subjects are to bear the adjectival property denoted by the stronger scalemate. We report the results of this experiment in Section 3.1. 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 3.2.

## 3.1 Experiment 3: Eliciting likelihoods

We experimentally measured the likelihood of the stronger scalar property obtaining with biased v. neutral subjects. 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.

### 3.1.1 Methods

**Participants** 62 native monolingual speakers of American English were recruited on Prolific and compensated approximately $2. One participant was removed due to failure to complete the experimental task (i.e. no responses were provided); data from 61 participants is reported below. The experiment was run on PCIbex.

**Materials and Procedures** 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*) v. neutral (*employees*) subject manipulation was tested within-participants. In addition to

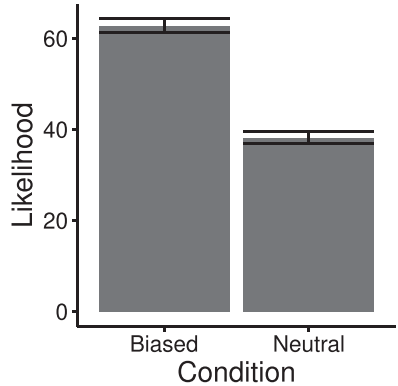


**Figure 1.** Mean likelihood (and 95% CI) of the biased v. neutral subject exhibiting the strong adjectival property.

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?*).

#### 3.1.2 Results and Discussion

Results are shown in Fig. 1. On average, biased subjects received higher ratings compared to neutral ones. A linear mixed effects regression model predicting likelihood ratings from the categorical predictor subject bias (biased v. neutral) was fitted to the data. The model was maximal, including random intercepts and slopes for participants and items, and the categorical predictor was scaled sum coded. Results show a significant effect of subject bias, such that biased subjects 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 subject nouns and provide us with a gradient as opposed to categorical likelihood measure.

### 3.2 Experiment 4: SI rates

#### 3.2.1 Methods

**Participants** 79 native monolingual speakers of American English were recruited on Prolific and compensated approximately $2.5. Four participants were excluded for having made more than four mistakes on the catch trials (this exclusion criteria was determined before data collection). Four further participants were excluded for taking more than 3 seconds to respond on critical trials, suggesting lack of attention (this exclusion criterion was determined upon examination of the data). Data from 71 participants is reported below. The experiment was administered on PCIbex.

**Materials and Procedure** Following van Tiel *et al.* (2016) (also Pankratz and 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*.

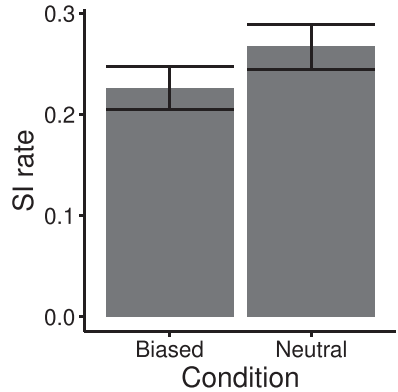


**Figure 2.** Mean by-subject SI calculation rate (and 95% CI) from the inference task.

The neutral v. biased subject manipulation was conducted within-participants. Since some scales share 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* → *not narrow*, *even* → *not odd*). Given that these items had an unambiguously correct answer ("Yes"), they were included to serve as catch trials.

### 3.2.2 Results and Discussion

Results are shown in Fig. 2. As shown in the plot, neutral subjects gave rise to higher SI rates compared to biased ones. We fitted a logistic mixed effects regression model to the data, predicting "Yes" v. "No" responses from the subject bias (biased v. 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 scaled sum coded. Model outputs confirm a significant effect of subject bias, such that neutral subjects led to significantly more SIs compared to biased ones ($\beta = 0.34$, $SE = 0.17$, $z = 2.0$, $p < 0.05$). This result was also replicated by a by-item analysis[8] (Fig. 3), where SI rates were regressed against the likelihood ratings obtained for each item in the likelihood experiment reported in Section 3.1. 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 in line with H1 in general—SI rates depend on the likelihood that the stronger adjective applies to an individual in the extension of the subject noun. More specifically, we found evidence for H1a, namely that SI is discouraged when the prior likelihood of the stronger statement is high, which for our experimental manipulation manifested as lower rates of SI calculation with biased subjects than with neutral ones.

Having seen that prior likelihoods affect SI rates, it is also worth considering their potential effect on another kind of pragmatic inference, namely negative strengthening. Negative strengthening is the phenomenon whereby an utterance such as *The employee is not brilliant* gets enriched from *The employee is less than brilliant* (its semantic meaning) to mean that the employee is in fact less than smart (Horn 1989, see i.a., Ruytenbeek *et al.* 2017, for experimental work). As has been noted in prior work, the task question of the

[8] Appendix I contains detailed by-item data for this experiment, as well as all subsequent experiments.

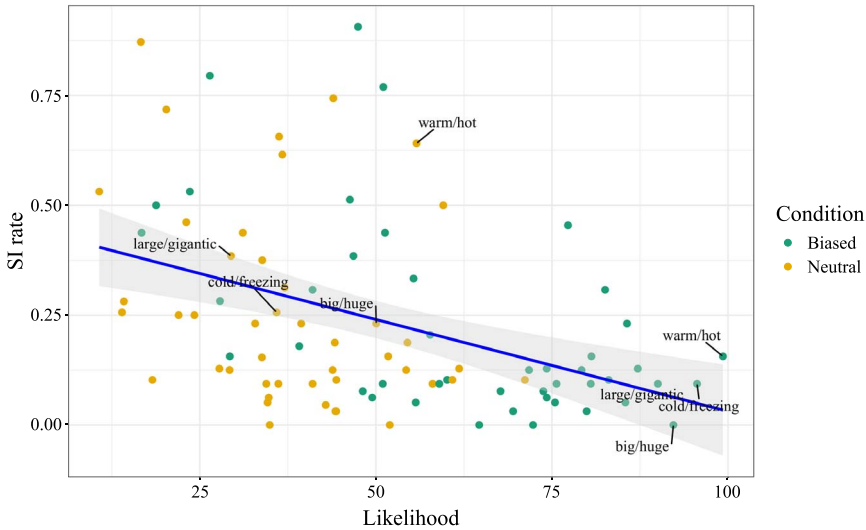


**Figure 3.** By-item correlation between SI rates and the likelihood of the subject noun exhibiting the stronger adjectival property.

inference task—used in much of the scalar diversity literature, including our Experiment 4, to probe SI rates—can lead to negative strengthening since it mentions the negated stronger alternative (Benz *et al.* 2018; Gotzner *et al.* 2018a,b). This might result in participants responding "No" to the "Would you conclude from this that... not brilliant?" question in the inference task irrespective of whether they calculated SI. In other words, the presence of negative strengthening can give the illusion of no SI calculation (Benz *et al.* 2018; Gotzner *et al.* 2018b). Turning now to the potential role of likelihoods, intuitively it seems that there may be differences in our items' propensity for negative strengthening depending on whether the neutral v. biased subject is used. Consider the following example:

(10) a) The employee is not brilliant.
b) The scientist is not brilliant.

Intuitively, (10-a) (neutral subject) is more likely to lead to negative strengthening, i.e. evoke the "The X is not smart" meaning, than (10-b) (biased subject). Suppose it is indeed the case that neutral subjects are more likely to lead to negative strengthening than biased ones; if negative strengthening gives the illusion of non-SI, then it is possible that neutral subjects lead to SI even more robustly than has been revealed by our experiment. That is to say, ruling out the potential confound of negative strengthening could strengthen the overall results already identified in our experiments: that neutral subjects lead to more SI than biased ones. Nonetheless, we leave an experimental exploration of the role of prior likelihoods in negative strengthening to future work.

## 4 Hypothesis 2: Threshold distance

In the previous section, we found evidence for Hypothesis 1a: that prior likelihood affects the robustness of SI such that biased subjects lead to less SI. Here, we turn to the question of whether this also means that neutral v. biased subjects differ in a known predictor of scalar diversity, semantic distance. Horn (1972, p. 112) notes that the further a stronger scalar term is on a scale from its weaker counterpart, the "safer" and "more likely to be justified"

it is to derive the SI (i.e. the negation of the stronger scalar). Horn (1972) demonstrates this intuition on the <*some*, *many*, *most*, *all*> scale: given an utterance containing *some*, he argues that *not all* is the strongest inference, while *not most* is weaker and *not many* is the weakest. In the context of scalar diversity, this idea of semantic distance has been experimentally tested by van Tiel *et al.* (2016). In their experiment, participants were presented with a weaker statement (e.g. *She is intelligent*) and its stronger alternative (*She is brilliant*) and had to rate on a Likert scale how much stronger the second statement was. The higher the average rating on the Likert scale, the larger the semantic distance between two scalar terms (e.g. <*smart*, *brilliant*>). The authors found a positive correlation between the results of this experiment and SI rates, which they take as evidence that the more distant a weak and a strong scalar term are on a scale, the more robust SI calculation is from that scale. This finding has since been replicated across a number of scalar diversity studies (Gotzner *et al.* 2018b; Pankratz and van Tiel 2021; Sun *et al.* 2018).

Here, we build on the semantic distance proposal by taking it to apply to adjectival thresholds. (Relatedly, see e.g. Gotzner *et al.* 2018a; Leffel *et al.* 2019, for the role of adjectival vagueness in SI calculation.) Crucially, we probe whether there is a difference between neutral and biased subjects in semantic distance. As mentioned, prior work has shown that smaller distances between weak and strong scalemates lead to lower rates of SI. Additionally, our previous experiments have shown that biased subjects also result in lower rates of SI. Given these two findings, it is possible that biased subjects correspond to reduced semantic distance compared to their neutral counterparts. This would amount to the finding that semantic distance not only varies across lexical scales but also within scales, and that this modulates SI rates. Importantly, we treat a potential change in distance between subject nouns as an empirical question. Specifically, we are interested in testing the following hypothesis:

(11) **Hypothesis 2 (H2)**: SI rates are modulated by the distance between the adjectival threshold of the two scalemates, such that greater distances encourage SI calculation, and this threshold distance varies based on subject noun.

Though our motivation for predicting that biased subjects lead to a reduction in semantic distance is primarily empirical, such a finding could be given the following theoretical explanation. With biased subjects, the adjectival threshold is raised for both weaker and stronger scalar terms, as compared to those scalar terms being predicated of neutral subjects. That is to say, the minimal threshold degree required of *brilliant-for-a-scientist* could be higher than *brilliant-for-an-employee*, and correspondingly the threshold for *smart-for-a-scientist* could also be higher than *smart-for-an-employee*. But it is conceivable that biased subjects in fact raise the threshold of the weak adjective more than they raise that of the strong adjective, such that a compression occurs at the top of the scale, ultimately leading to a smaller distance between the weak and strong thresholds.[9] In particular, this is possible with so-called bounded scales (Horn 1972; van Tiel *et al.* 2016), where the stronger alternative denotes the endpoint of a scale (e.g. <*difficult*, *impossible*>). With such scales, since it is impossible to raise the threshold of the stronger adjective, only the threshold of the weaker adjective may be raised with biased subjects, creating a smaller distance. Additionally, we argue that a similar compression could occur with scales where the stronger term is an extreme adjective. Based on observations such as extreme adjectives

[9] Prior work has also revealed broadly similar effects of context modulating adjectival thresholds and hence semantic distance. For instance, Alexandropoulou *et al.* (2022) have shown that context, operationalized in their study as the presence of a contrast object, affects SI calculation from relative adjectives. The authors argue that what may underlie this finding is that the threshold for the weak relative adjective is lowered when a contrast object is present, and this in turn increases the weak adjective's semantic distance from its stronger alternative, ultimately impacting SI calculation.

resisting intensification (e.g. ∗*very gigantic/gorgeous*) and allowing for modification by degree modifiers that target the upper bound of the relevant adjectival scale (e.g. *completely gorgeous/gigantic*), it has been proposed that extreme adjectives are upper-bounded (Paradis 2001). Alternatively, extreme adjectives have been analysed as denoting extreme degrees that are beyond the set of contextually relevant degrees (Morzycki 2012). Extreme degrees on this analysis are taken to be undifferentiated in the context, with individuals bearing extreme degrees mapped to the same equivalence class (Cresswell 1976). That is to say, with extreme adjectives, it is possible that though the threshold is higher for *brilliant-for-a-scientist* than it is for *brilliant-for-an-employee*, the two are hard to distinguish, as the difference between extreme degrees is not relevant in the context of use. Thus, with both bounded and extreme scales, attempting to raise the adjectival threshold for both the weaker and stronger scalar with biased subjects might amount to only raising the weaker threshold (or raising the weaker threshold to a greater extent), ultimately resulting in a compression of thresholds and a smaller semantic distance. Crucially, of the 45 scales tested in our experiments, 41 are either bounded or extreme.[10]

More generally, the exploration of H2 also serves a methodological purpose. Our previous findings revealed that sentential context has an impact on the likelihood of SI calculation, and specifically that a by-subject difference in SI rates arises. Given this, it is worth confirming whether a prior correlate of scalar diversity, namely semantic distance, holds across the board or whether this predictor is also sensitive to sentential context in the same way that SI rates are. In order to evaluate H2, we conduct an experiment with the goal of obtaining degree priors for the same scales and subject nouns tested in our previous experiments. We then use the elicited degrees to construct a distance metric, to be more precisely defined in (14), and examine whether the distance metric is a predictor of SI rates, and whether it varies between subject nouns.

## 4.1 Experiment 5: Eliciting degree priors

### 4.1.1 Methods

**Participants** 60 native monolingual speakers of American English participated in each of the two strong adjective conditions and were compensated $1.60 or $2 (depending on time). One participant was excluded for not completing the task (i.e. the participant did not provide any responses). A total of 120 participants were recruited for each of the two weak adjective conditions, since we anticipated having to discard data where SI was calculated. These participants were compensated $2.40 or $3 depending on time. Lastly, to supplement data collection in the weak conditions (see fn. 13), an additional 60 participants (native monolingual speakers of American English) were recruited for each of the biased and neutral subject conditions. They saw a small subset of items and were compensated $0.60. Data from three participants was not included in subsequent analyses due to data loss. Participant recruitment took place on Prolific and the experiment was run on PCIbex.

**Materials and Procedures** An experiment was conducted to obtain prior distributions regarding the degree to which the individual denoted by the subject noun bears the adjectival property. We use degree estimates as a proxy for thresholds, since participants presumably do not have intuitions about thresholds (e.g. the degree of intelligence at which someone counts as *smart*). As mentioned in Section 1.2, we take it that higher thresholds elicit

[10] We categorized each scale as bounded v. non-bounded based on prior work, adopting classifications from Gotzner *et al.* (2018b), Pankratz and van Tiel (2021), and Ronai (2022). For extremeness, we adopted classifications from Gotzner *et al.* (2018b) and Pankratz and van Tiel (2021); this left <*understandable*, *articulate*> unclassified, which we classify as non-extreme based on the diagnostic that extreme adjectives should be modifiable by *downright* and *flat-out*. Altogether, only 4 out of our 45 scales, namely <*cool*, *cold*>, <*warm*, *hot*>, <*wary*, *scared*>, and <*understandable*, *articulate*>, are neither bounded nor extreme. See Appendix I for all classifications.

higher degrees on average. Therefore, the collected degree estimates should reflect changes in thresholds that may be induced by different subject nouns via their effect on CCs.

We begin by describing the conditions probing such degree distributions for the stronger adjectives. Participants were presented with an utterance from (12), either in the neutral (12-a) or biased (12-b) condition.

(12) a) The employee is brilliant. neutral, strong
b) The scientist is brilliant. biased, strong

On the same screen, participants also saw a sliding scale with endpoints labeled "0" and "100", and they were asked the question "On a 0–100 scale, how smart is the {scientist/employee}?". They provided their judgment by picking a point on the sliding scale. The task questions ("On a 0–100 scale...") always relied on the weaker term from the scale, i.e. *smart* for *brilliant*.

For the weaker adjectives (e.g. *smart*), we wanted to rule out the possibility that participants would calculate the SI from *The employee is smart* and provide degree estimates given an enriched *The employee is smart, but not brilliant* meaning. In order to do this, we combined the sliding scale task with the inference task. Participants first saw an utterance containing the weaker scalar term in the neutral (13-a) or biased (13-b) condition.

(13) a) The employee is smart. neutral, weak
b) The scientist is smart. biased, weak

These utterances were paired with the same sliding scale and task question ("On a 0–100 scale...") as the stronger adjectives, and participants responded by picking a point on the sliding scale. Crucially, on each trial, the sliding scale task was followed by an inference task, which probed whether participants had calculated the SI. That is, after an utterance from (13), the next screen presented the same utterance again (Mary: *The {employee, scientist} is smart*), and participants had to answer the question "Would you conclude from this that Mary thinks the {employee, scientist} is not brilliant?".[11] If a participant answered "Yes", we took that to mean that they had calculated the *smart but not brilliant* SI. Consequently, we removed their response on the sliding scale task from the analysis. If they answered "No", we took that to mean that they had provided a degree estimate for the non-SI-enriched meaning of *smart*. Degree estimates that were followed by a "No" response on the inference task were retained for analysis.[12]

The weak v. strong and neutral v. biased manipulations were conducted between-participants. In addition to the 45 critical items, the experiment included 3 practice and 5 fillers items.[13] 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.

[11] Similarly to the main SI experiment (Section 3.2), we made sure that no participant made an SI judgment on *not brilliant* or *not delicious* twice.

[12] Note that we are treating SI calculation here as a binary possibility. This is in contrast to theories that take the availability of an SI-enriched reading to be gradable (Franke and Jäger 2016; Goodman and Frank 2016). Without aiming to adjudicate among these competing conceptualizations of SI, we note that participants' degree response was indeed different depending on their response in the inference task—see Appendix II. This suggests that employing a binary filter is useful in eliminating at least some of the influence of SI calculation on degree estimates.

[13] In the weak adjective condition, how much data we were able to analyse depended on the rate of SI calculation from a given scale: the higher the SI rate of a scale, the more data we had to discard in order to rule out the possibility that degree estimates were given based on SI-enriched meanings. To supplement data collection for high SI rate scales, we conducted an additional experiment that tested only a subset of scales: 7 for the biased and 9 for the neutral condition. In addition to the 7/9 critical items, these experiments included 2 practice and 3 fillers items.

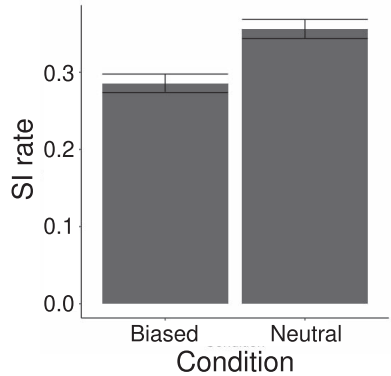


**Figure 4.** Mean by-subject SI calculation rate (and 95% CI) from the inference task conducted as part of the degree elicitation experiment.

### 4.1.2 Results and Discussion

We begin by analysing the results of the weak conditions, which used the combined task. The inference task's results replicated what we found in the standalone inference task (Section 3.2): neutral subjects led to SI more robustly than biased ones. These results are shown in Fig. 4 (compare Fig. 2); the same logistic mixed effects model fitted to the data pertaining to the standalone SI rates experiment was fit to the current results, replicating previous findings: neutral subjects gave rise to significantly higher SI rates compared to biased subjects ($\beta = 0.46$, SE= 0.086, $z = 5.4$, $p < 0.001$). Recall that the combined task was conducted so we could eliminate degree responses that may have been enriched with SI. In what follows, we therefore exclude from analysis sliding scale data from the weak conditions that was followed by a "Yes" response in the inference task.[14]

Taking all four data sets together (weak neutral, weak biased, strong neutral, strong biased), we find that on average, the mean ratings were higher for strong scalemates compared to weak scalemates for both neutral and biased subjects. Ratings were also higher for biased than for neutral subjects. This was confirmed by a linear mixed effects model predicting degree responses from SUBJECT BIAS, STRENGTH and their interaction, with by-condition random slopes and by-item random intercepts. Both categorical predictors were scaled sum coded. The main effects were significant (SUBJECT BIAS: $\beta = -1.36$, SE= 0.35, $t = -3.86$, $p < 0.0001$; STRENGTH: $\beta = -5.30$, SE= 0.65, $t = -8.1$, $p < 0.0001$) but the interaction was not ($\beta = -0.32$, SE= 0.71, $t = -0.44$, $p > 0.6$).

We now turn to evaluating H2. To determine the effect of threshold distance on SI rates, we computed Cohen's $d$ for each of the 45 scales using the elicited degree estimates. This metric allowed us to quantify the distance between the thresholds of two scalemates on a shared adjectival scale in the biased and the neutral conditions respectively. As seen in equation (14), the distance $d_n$ was computed by subtracting the mean $\mu$ of the degree estimates for the weak scalemate $w_n$ from the strong scalemate $s_n$, where $n$ stands for a particular subject noun. The difference between means was subsequently standardized by dividing by the pooled standard deviation, i.e. the average of the variances of the relevant random variables.

[14] After exclusions, we still mostly retain as many data points per scale as had been collected for the strong conditions. This is thanks to running double the number of participants in the weak conditions, and supplementing data collection for the highest SI rate scales—see fn. 13. In each of the strong conditions (biased and neutral), each scale has 60 observations; in the weak conditions, we meet or exceed this number in 36/45 scales in the biased and 37/45 scales in the neutral condition.

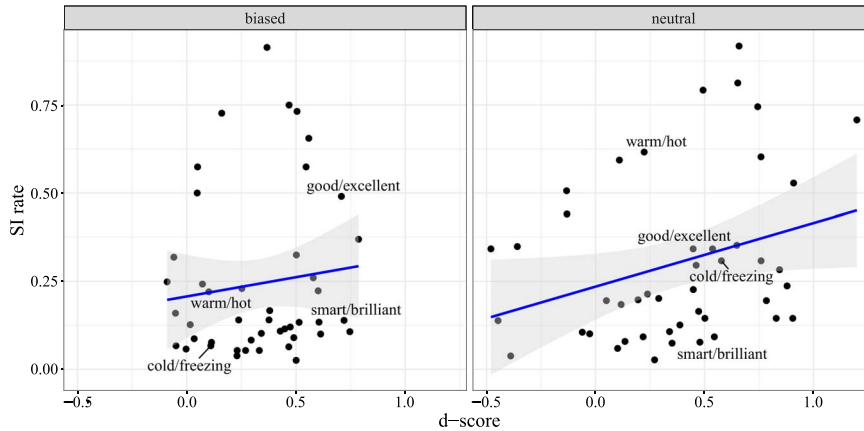


**Figure 5.** By-item correlation between SI rates (Experiments 4–5) and $d_n$ scores (Experiment 5), which index adjectival threshold distance, for biased (left) and neutral (right) subject nouns.

$$(14)\ d_n = \frac{\mu_{s_n} - \mu_{w_n}}{\sqrt{\frac{\sigma^2_{s_n} + \sigma^2_{w_n}}{2}}}$$

Numerically, $d_n$ scores with biased subjects were on average lower ($M = 0.33$) than $d_n$ scores with neutral subjects ($M = 0.37$). However, a paired t-test revealed that this difference failed to reach significance ($t(44) = 0.86, p > 0.3$).[15] Next, to validate our distance metric $d_n$, we sought to replicate effects of semantic distance found by van Tiel *et al.* (2016) and subsequent work. We first correlated the $d_n$ scores with the SI rates from Experiment 4 and found a positive relationship that nonetheless failed to reach significance ($r = 0.2$, $p < 0.062$). We reasoned that our failure to replicate the significant correlation between SI rates and semantic distance could be due to insufficient power. For this reason, we pooled the inference-task-based SI calculation data from both Experiment 4 and Experiment 5, and correlated the SI rates obtained this way with the $d_n$ scores.[16] Here, we successfully replicate the semantic distance effect, finding significantly higher SI rates with scales where the two scalemates have more distant adjectival thresholds ($r = 0.23, p < 0.03$). Subsequent analyses therefore use pooled SI data.[17]

[15] Further visual inspection of the data (see x axis of Fig. 5) suggested that there was less variance among $d_n$ scores with biased subjects than with neutral subjects. Levene's test confirms that this difference in variance is statistically significant ($F(1, 88) = 6.30, p < 0.02$). However, in light of the non-significant difference in $d_n$ scores across different subject nouns (see paired t-test above), we do not wish to attribute greater theoretical importance to the difference in variance.

[16] To increase our confidence in the validity of analysing the pooled data, we also conducted a mini meta-analysis on the results of the two separate studies for the effect of $d_n$ scores on SI rates, following the procedure recommended by Goh *et al.* (2016). We used fixed effects in which the mean effect size was first obtained (i.e. the mean correlation coefficient). All correlations were Fisher's $z$ transformed for analyses and converted back to Pearson correlations for ease of presentation; $p$-values were obtained using Stouffer's $Z$-test (all tests were two-tailed). Results show that $d_n$ scores were positively correlated with SI rates ($Mr = 0.21$, $p = 0.05$). The full list of coefficients and associated statistics is as follows: Experiment 4 $r = 0.198$; Experiment 5 $r = 0.214$; $Mr_z = 0.209$; $Mr = 0.206$; combined $Z = 1.955$, where $Mr_z$ corresponds to the weighted mean correlation (Fisher's $z$ transformed) and $Mr$ corresponds to the weighted mean correlation converted from $r_z$ to $r$. The results from our mini meta-analysis therefore suggest that the effect sizes of the two studies (Experiment 4 and Experiment 5) were highly comparable to one another, as well as to the estimates obtained with the pooled Experiment 4–5 data.

[17] We checked whether the previously reported likelihood effect (Experiment 4) is replicated using SI rates from both Experiment 4 and 5, and found that it is ($r = -0.47, p < 0.001$).

We now tackle the question of whether the semantic distance-SI rate correlation is reliable across sentential contexts. Figure 5 plots $d_n$ scores against SI rates in the neutral and biased conditions respectively. Visual inspection suggests that the positive relationship between threshold distance and SI rates seems to be more pronounced for biased nouns. In order to probe whether $d_n$ scores' ability to predict SI rates is indeed affected by sentential context, we fit a linear regression model predicting SI rates from $d_n$ score, SUBJECT BIAS (biased v. neutral, scaled sum coded) and their interaction. Crucially, the model revealed no significant interaction ($\beta = 0.07$, $SE = 0.16$, $t = 0.43$, $p > 0.6$). This suggests that—despite the slight difference suggested by Fig. 5—the positive correlation between SI rates and threshold distance is not significantly modulated by subject nouns.

Altogether, our findings do not reveal evidence that the semantic distance between a weak adjective and its stronger counterpart is modulated by whether the adjectives are predicated of biased or neutral nouns. Nor do we find evidence that the ability of semantic distance to predict scalar diversity would depend on the subject noun. Of course, null results—such as the lack of evidence for an impact of subject nouns on semantic distance reported here—should always be interpreted with caution. In particular, it is possible that there exists a modulating effect of subject nouns, but our experiments failed to detect it. One reason this might be is the possibility that when providing threshold estimates in Experiment 5, participants judged adjectives relative to the comparison class invoked by the subject noun that the adjective was predicated of, given the task questions "On a 0–100 scale, how smart is the scientist?" and "On a 0–100 scale, how smart is the employee?".[18] That is, it may have been the case that rather than judging where a *smart scientist* or *smart employee* would fall on a general scale of smartness, participants were evaluating where a *smart scientist* would fall on a smartness scale for scientists and where a *smart employee* would fall on a smartness scale for employees. One finding that speaks against this possibility is the main effect of subject bias: we found overall higher thresholds for biased subjects than for neutral ones, which is unexpected if each sentence was evaluated relative to its invoked comparison class. Nonetheless, a variant of Experiment 5 with a task question like "On a 0–100 scale, how smart of a person is the {scientist/employee}?" might more strongly mitigate against the possibility that the threshold estimates given are influenced by the invoked comparison class, and hence such an experiment might be able to isolate an effect of subject nouns on semantic distance—we leave this investigation to future work.

## 5 General Discussion

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 subjects the adjectives were predicated of. For each scale, two subjects were established: a biased one, 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 one, where this likelihood was not especially high (e.g. *employee* for <*smart*, *brilliant*>). We found a significant effect of the biased v. neutral subject manipulation across the board: neutral subjects led to higher rates of SI calculation.

Our finding that sentential contexts introduce within-scale variation in SI rates is expected given previous work that tested SI calculation from naturally occurring utterances extracted from corpora (Degen 2015; Sun *et al.* 2023). But it seemingly goes against van Tiel *et al.* (2016)'s original scalar diversity study, which found no difference across carrier sentences

[18] We thank an anonymous reviewer for suggesting this possibility.

within the same scale. At the same time, a number of key differences between our work and van Tiel *et al.* (2016)'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, we specifically focused on making one carrier sentence per scale "biased", while van Tiel *et al.* (2016)'s method of eliciting these sentences merely asked participants for a natural-sounding completion, which likely gave more neutral results overall. These differences might explain why van Tiel *et al.* (2016)'s investigation of within-scale variation produced a null result, while we did find a significant effect of different sentential contexts.

The main hypothesis (H1) that we explored for the effect of sentential context on within-scale SI rate variation concerned prior likelihood; according to H1, SI calculation is directly affected by how likely it is that the stronger scalar property obtains. It has been argued that high prior likelihood of the stronger alternative discourages SI calculation (H1a): if a hearer knows that scientists are likely to be brilliant, she would be less inclined to derive the *Scientists are smart but not brilliant* SI. Existing work has revealed somewhat limited empirical evidence for H1a, with Degen *et al.* (2015) finding only a small effect of prior likelihood on SI calculation for the <*some*, *all*> scale, and Tsvilodub *et al.* (2023) finding that such likelihood effects do not extend to the <*or*, *and*> scale—motivating further exploration of this hypothesis with more context-sensitive adjectival scales. On the other hand, we have also argued that the opposite prediction to H1a can be made (which we called H1b). Since SI arises from the non-utterance of a stronger alternative, hearers could attribute especially great importance to the speaker's choice to use the less informative weaker scalemate in cases where the stronger one was a priori likely to be true, which would then lead to higher rates of SI. In terms of our experimental manipulation, these two potential effects of likelihood would manifest as biased subjects leading to less SI than neutral ones under H1a, and biased subjects leading to more SI under H1b. Our experimental findings were in line with H1a, providing further evidence for the role of priors in SI calculation.

In a separate set of experiments, we also tested the modulating effect of adjectival threshold distance between the weaker v. stronger scalemate on SI rates. This was motivated by existing literature (i.a. Gotzner *et al.* 2018b; Pankratz and van Tiel 2021; Sun *et al.* 2018; van Tiel *et al.* 2016) finding that smaller semantic distance corresponds to lower SI rates, and our previous experiments finding lower SI rates with biased subjects—raising the possibility that semantic distance varies not only across scales, but also within scales, with biased subjects showing smaller distances than neutral ones. We argued that if biased subjects indeed lead to smaller semantic distance, this may be because they raise the threshold of adjectives compared to neutral subjects, but they might raise the weak adjective's threshold more than the strong adjective's—a possibility for bounded and extreme scales. Ultimately, our experiments did not reveal evidence for such a difference in semantic distance between subject nouns, nor did we find that the previously identified positive correlation between SI rates and semantic distance would vary depending on the subject noun. Instead, semantic distance continued to be a predictor of scalar diversity irrespective of sentential context. At the same time, we noted that our choice of task question in the threshold elicitation experiment may have obscured the effect of subject nouns on semantic distance.

Lastly, let us touch on two further methodological considerations 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 (Section 2.1) and what cutoff to employ for counting a scale as having "passed" 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 affects likelihood of SI for a large number of different scales—has for those previous scalar diversity studies that tested only a single carrier sentence per scale. In principle, it is possible that the uncontrolled effect of sentential contexts had introduced a confound in such prior work. But for this to necessitate the reinterpretation of previous findings, the following would need to have obtained: a systematic bias in previous experiments, such that some scales had been tested with what would count as a neutral subject, and some others with what would count as a biased subject. 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 subject, then the difference in SI rates could have arisen as an artifact of the sentence frames. Two existing findings speak against this possibility. First, scale-intrinsic factors such as boundedness have been shown to reliably predict scalar diversity in different experiments, suggesting that at least some of the inter-scale variation is tied to properties of lexical scales. Second, Sun *et al.* (2023) have shown that when lexical scales are ordered according to their SI rate, their relative order largely remains the same no matter whether they are tested with two (Sun *et al.* 2018), three (van Tiel *et al.* 2016), or fifty (Sun *et al.* 2023) carrier sentences per scale.

But while we are not suggesting that across-scale variation in SI rates could be reduced to an illusion arising from carrier sentences, it remains the case that our experiments revealed significant within-scale variation. Future work should therefore pay closer attention to controlling carrier sentences, and testing a larger variety of them. This will also help us gain a fuller understanding of how much variation in SI calculation can be attributed to the identity of lexical scales v. contextual cues—for similar arguments, see also Degen (2021).

## 6 Conclusion

This paper investigated variation in the robustness of SI calculation both within- and across-different lexical scales. In other words, we tested the role of different sentential contexts in scalar diversity. Focusing on lexical scales formed by two gradable adjectives, we showed that the rate of SI calculation is significantly different based on the choice of the subject noun. Specifically, we found that if the prior likelihood of the stronger alternative is high with a particular noun (e.g. *brilliant scientist*), SI calculation is discouraged as compared to more neutral subject nouns. We additionally explored the interaction of subject nouns with semantic distance (Horn 1972; van Tiel *et al.* 2016), which we operationalized as adjectival threshold distance (between e.g. *smart* and *brilliant*), but we did not find evidence that semantic distance or its ability to predict SI rates varies across subject nouns. Altogether, the findings reported in this paper have methodological consequences: future work should pay attention to properties of carrier sentences in the testing of pragmatic inferences. Our work is also informative for theories of SI calculation and variation therein, providing more robust evidence for the role of priors (Degen *et al.* 2015; Tsvilodub *et al.* 2023) from a large variety of different scales.

## Acknowledgements

We thank the editor David Barner, two anonymous reviewers, Julian Grove, Michael Tabatowski, and the audience at SALT 33 for their invaluable input. This material is partially based upon work supported by the National Science Foundation under Grant No. #BCS-2041312.

## Conflict of interest

The authors declare no conflict of interests.

## Data availability

Stimuli, data, and the scripts used for data visualization and analysis can be found in the following OSF repository: https://osf.io/a2gje/?view_only=edc20e113476422d86698688106891cc. The studies reported in this paper were not pre-registered.

## References

Alexandropoulou, S. *et al.* (2022) 'Incremental pragmatic interpretation of gradable adjectives: The role of standards of comparison'. *Semantics and Linguistic Theory*, 1: 481–97.

Baker, R. *et al.* (2009) 'On the non-unified nature of scalar Implicature: An empirical investigation'. *International Review of Pragmatics*, 1: 211–48.

Barner, D. and Snedeker, J. (2008) 'Compositionality and statistics in adjective acquisition: 4-year-olds interpret tall and short based on the size distributions of novel noun referents'. *Child Development*, 79: 594–608.

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

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

Cremers, A. (2022) 'Interpreting gradable adjectives: Rational reasoning or simple heuristics'. *Empirical Issues in Syntax and Semantics*, 14: 31–61.

Cresswell, M. J. (1976) 'The semantics of degree'. In B. Partee (ed.), *Montague Grammar*. 261–92.

de Marneffe, M.-C. and Tonhauser, J. (2019) 'Inferring meaning from indirect answers to polar questions: The contribution of the rise-fall-rise contour'. In M. Zimmermann, K. von Heusinger and E. Onea (eds), *Current Research in the Semantics/Pragmatics Interface, volume 36, Questions in Discourse*. The Netherlands. Brill, Leiden. 132–63.

Degen, J. (2013) 'Alternatives in Pragmatic Reasoning', PhD thesis, University of Rochester.

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

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

Degen, J., Tessler, M. H. and Goodman, N. D. (2015) 'Wonky worlds: Listeners revise world knowledge when utterances are odd'. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*. 548–53.

Doran, R. *et al.* (2012) 'A novel experimental paradigm for distinguishing between what is said and what is impli-cated'. *Language*, 88: 124–54.

Ebeling, K. S. and Gelman, S. A. (1988) 'Coordination of size standards by young children'. *Child Development*, 888–96.

Ebeling, K. S. and Gelman, S. A. (1994) 'Children's use of context in interpreting "big" and "little"'. *Child Development*, 65: 1178–92.

Fairchild, S. and Papafragou, A. (2021) 'The role of executive function and theory of mind in pragmatic computations'. *Cognitive Science*, 45: e12938.

Foppolo, F. (2013) 'Do children know when their room counts as clean?' In *NELS 40: Proceedings of the 40th Annual Meeting of the North East Linguistic Society: Volume One (Volume 1)*. GLSA. 205–18.

Frank, M. C. and Goodman, N. D. (2012) 'Predicting pragmatic reasoning in language games'. *Science*, 336: 998–8.

Franke, M. and Jäger, G. (2016) 'Probabilistic pragmatics, or why bayes' rule is probably important for pragmatics'. *Zeitschrift für Sprachwissenschaft*, 35: 3–44.

Gelman, S. A. and Ebeling, K. S. (1989) 'Children's use of nonegocentric standards in judgments of functional size'. *Child Development*, 920–32.

Geurts, B. (2010) *Quantity Implicatures*. Cambridge University Press.

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

Goh, J. X., Hall, J. A. and Rosenthal, R. (2016) 'Mini meta-analysis of your own studies: Some arguments on why and a primer on how'. *Social and Personality Psychology Compass*, 10: 535–49.

Goodman, N. D. and Frank, M. C. (2016) 'Pragmatic language interpretation as probabilistic inference'. *Trends in Cognitive Sciences*, 20: 818–29.
Gotowski, M. and Syrett, K. (2020) 'Investigating the hypothesis space of children's interpretation of comparatives'. In *Proceedings of the 44th Annual Boston University. Conference on Language Development*. Cascadilla Press. 154–67.
Gotzner, N., Solt, S. and Benz, A. (2018a) 'Adjectival scales and three types of implicature'. *Semantics and Linguistic Theory*, 28: 409.
Gotzner, N., Solt, S. and Benz, A. (2018b) '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*. Harvard University Press. 41–58.
Heim, Irene (2000) 'Degree operators and scope'. In *Pages 40–64 of: Semantics and Linguistic Theory (SALT) 10*.
Horn, L. R. (1972) 'On the Semantic Properties of Logical Operators in English'. PhD thesis, UCLA.
Horn, L. R. (1989) *A Natural History of Negation*. University of Chicago Press. Chicago.
Hu, J. *et al.* (2023) 'Expectations over unspoken alternatives predict pragmatic inferences'. *Association for Computational Linguistics*, 11: 885–901.
Hu, J., Levy, R. and Schuster, S. (2022) 'Predicting scalar diversity with context-driven uncertainty over alternatives'. *Workshop on Cognitive Modeling and Computational Linguistics*, 68–74.
Jasbi, M., Waldon, B. and Degen, J. (2019) 'Linking hypothesis and number of response options modulate inferred scalar implicature rate'. *Frontiers in Psychology*, 10.
Kennedy, C. (2007) 'Vagueness and grammar: The semantics of relative and absolute gradable adjectives'. *Linguistics and Philosophy*, 30: 1–45.
Kennedy, C. and McNally, L. (2005) 'Scale structure, degree modification, and the semantics of gradable predicates'. *Language*, 81: 345–81.
Kennedy, C. (1999) *Projecting the Adjective: The Syntax and Semantics of Gradability and Comparison*. Garland.
Lassiter, D. and Goodman, N. D. (2013) 'Context, scale structure, and statistics in the interpretation of positive-form adjectives'. In *Semantics and Linguistic Theory (SALT 23)*. 587–610.
Leffel, T. *et al.* (2019) 'Vagueness in Implicature: The case of modified adjectives'. *Journal of Semantics*, 36: 317–48.
Li, E., Schuster, S. and Degen, J. (2021) 'Predicting scalar inferences from "or" to "not both" using neural sentence encoders'. In *Proceedings of the Society for Computation in Linguistics 2021*, 446–50.
Morzycki, M. (2012) 'Adjectival extremeness: Degree modification and contextually restricted scales'. *Natural Language & Linguistic Theory*, 30: 567–609.
Pankratz, E. and van Tiel, B. (2021) 'The role of relevance for scalar diversity: A usage-based approach'. *Language and Cognition*, 13: 562–94.
Paradis, C. (2001) 'Adjectives and boundedness'. *Cognitive Linguistics*, 12: 47–65.
Ronai, E. (2022) 'Scales, Alternatives, Context: Experimental Investigations Into Scalar Inference', PhD thesis, The University of Chicago.
Ronai, E. and Xiang, M. (2020) 'Pragmatic inferences are QUD-sensitive: An experimental study'. *Journal of Linguistics*, 1–30.
Ronai, E. and Xiang, M. (2021) 'Exploring the connection between question under discussion and scalar diversity'. *Linguistic Society of America (LSA)*, 6: 649–62.
Ronai, E. and Xiang, M. (2022) 'Three factors in explaining scalar diversity'. *Sinn und Bedeutung*, 26: 716–33.
Ruytenbeek, N., Verheyen, S. and Spector, B. (2017) 'Asymmetric inference towards the antonym: Experiments into the polarity and morphology of negated adjectives'. *Glossa: a journal of general linguistics*, 2: 92.
Solt, S. and Gotzner, N. (2012a) 'Experimenting with degree'. In *Semantics and Linguistic Theory*, pp. 166–87.
Solt, S. and Gotzner, N. (2012b) 'Who here is tall? Comparison classes, standards and scales'. In *Pages 79–83 of: International Conference on Linguistic Evidence*.
Sun, C., Tian, Y. and Breheny, R. (2018) 'A link between local enrichment and scalar diversity'. *Frontiers in Psychology*, 9.

Sun, C., Tian, Y. and Breheny, R. (2023) 'A corpus-based examination of scalar diversity'. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.
Syrett, K., Kennedy, C. and Lidz, J. (2010) 'Meaning and context in children's understanding of gradable adjectives'. *Journal of Semantics*, 27: 1–35.
Syrett, K., Kennedy, C. and Lidz, J. (2009) 'Meaning and context in children's understanding of gradable adjectives'. *Journal of Semantics*, 27: 1–35.
Tessler, M. H. and Goodman, N. D. (2022) 'Warm (for winter): Inferring comparison classes in communication'. *Cognitive Science*, 46: e13095.
Tessler, M. H. *et al.* (2020) 'Informational goals, sentence structure, and comparison class inference'. In *Proceedings of the Annual Conference of the Cognitive Science Society*.
Tsvilodub, P., van Tiel, B. V. and Franke, M. (2023) 'The role of relevance, competence, and priors for scalar inferences'. *Proceedings of Experiments in Linguistic Meaning (ELM)*, 2: 288–98.
van Tiel, B. *et al.* (2016) 'Scalar diversity'. *Journal of Semantics*, 33: 137–75.
von Stechow, A. (1984) 'Comparing semantic theories of comparison'. *Journal of Semantics*, 3: 1–77.
Westera, M. and Boleda, G. (2020) 'A closer look at scalar diversity using contextualized semantic similarity'. *Sinn und Bedeutung*, 24: 439–54.
Zehr, J. and Schwarz, F. (2018) 'PennController for internet based Experiments (IBEX)'. https://doi.org/10.17605/OSF.IO/MD832.
Zondervan, A., Meroni, L. and Gualmini, A. (2008) 'Experiments on the role of the question under discussion for ambiguity resolution and implicature computation in adults'. In T. Friedman and S. Ito (eds), *Proceedings of Semantics and Linguistic Theory (SALT) 18*. 765–77.

## Appendix I

Scales excluded on the basis of the norming studies (Experiment 1):

- Excluded based on the symmetric entailment ("*but not*") test: <*annoyed*, *angry* >, <*busy*, *full* >, <*casual*, *sloppy* >, <*grey*, *black* >, <*memorable*, *unforgettable* >, <*rough*, *unfriendly* >, <*silly*, *ridiculous* >, <*thin*, *skinny* >
- Excluded based on the cancellability ("*even*") test: <*adequate*, *good* >, <*annoyed*, *angry* >, <*busy*, *full* >, <*calm*, *unflappable* >, <*chubby*, *fat* >, <*comfortable*, *luxurious* >, <*content*, *happy* >, <*cute*, *beautiful* >, <*enjoyable*, *great* >, <*grey*, *black* >, <*honest*, *blunt* >, <*intelligent*, *brilliant* >, <*likely*, *certain* >, <*low*, *depleted* >, <*mediocre*, *bad* >, <*nice*, *great* >, <*polite*, *friendly* >, <*rare*, *extinct* >, <*rough*, *unfriendly* >, <*satisfactory*, *impeccable* >, <*scarce*, *unavailable* >, <*silly*, *idiotic* >, <*silly*, *ridiculous* >, <*snug*, *tight* >, <*thin*, *skinny* >, <*unkind*, *nasty* >, <*unsettling*, *horrific* >

**Table 1.** Likelihood (Experiment 3).

| Scale | Likelihood Neutral | Likelihood Biased |
|---|---|---|
| attractive/stunning | 34.43556 | 75.42800 |
| big/huge | 50.04600 | 92.24139 |
| big/enormous | 58.04306 | 87.18400 |
| bright/brilliant | 42.87320 | 64.66528 |
| calm/meditative | 29.22222 | 74.25440 |
| cheap/free | 16.62040 | 47.45889 |
| cold/frosty | 44.16361 | 85.42480 |
| cold/freezing | 35.89720 | 95.59722 |
| cool/cold | 51.74194 | 82.54160 |
| dark/black | 13.94840 | 80.60250 |
| difficult/impossible | 10.71694 | 26.41720 |
| fat/obese | 44.37800 | 50.97583 |
| funny/hilarious | 36.14833 | 60.10520 |
| good/perfect | 20.23600 | 23.58250 |
| good/excellent | 37.04361 | 46.30600 |
| happy/ecstatic | 18.26720 | 75.64278 |
| happy/delighted | 44.35528 | 55.65480 |
| hard/unsolvable | 36.71640 | 16.72389 |
| hot/boiling | 33.86361 | 27.86400 |
| hot/scalding | 61.82080 | 29.25722 |
| hungry/starving | 24.22583 | 67.69440 |
| large/gigantic | 29.43320 | 90.03778 |
| loud/deafening | 54.47417 | 85.65360 |
| old/ancient | 32.85480 | 79.20306 |
| palatable/delicious | 59.56611 | 77.27560 |
| poor/destitute | 33.81520 | 79.92472 |
| pretty/beautiful | 44.24972 | 83.05480 |
| pretty/gorgeous | 71.15520 | 74.26194 |
| quiet/silent | 43.85778 | 41.00320 |
| quiet/inaudible | 23.07360 | 51.29361 |
| red/scarlet | 14.21639 | 55.34560 |
| scared/petrified | 27.77160 | 49.47889 |
| sleepy/asleep | 36.24889 | 51.05360 |
| small/tiny | 39.39320 | 80.50861 |
| smart/brilliant | 34.92333 | 71.74000 |
| soft/mushy | 31.10160 | 57.70222 |
| special/unique | 34.62361 | 59.00160 |
| tasty/delicious | 51.97080 | 72.31278 |
| thick/impenetrable | 21.98333 | 46.80880 |
| tough/impossible | 43.95120 | 18.79056 |
| ugly/hideous | 34.78417 | 73.77600 |
| understandable/articulate | 60.86280 | 69.48417 |
| unhappy/miserable | 41.01444 | 48.13520 |
| warm/hot | 55.76680 | 99.27444 |
| wary/scared | 54.29444 | 39.11680 |

**Table 2.** Experiment 4 SI rates.

| Scale | Exp4 SI Neutral | Exp4 SI Biased |
|---|---|---|
| attractive/stunning | 0.0937500 | 0.0512821 |
| big/huge | 0.2307692 | 0.0000000 |
| big/enormous | 0.0937500 | 0.1282051 |
| bright/brilliant | 0.0454545 | 0.0000000 |
| calm/meditative | 0.1250000 | 0.1282051 |
| cheap/free | 0.8717949 | 0.9062500 |
| cold/frosty | 0.1875000 | 0.0512821 |
| cold/freezing | 0.2564103 | 0.0937500 |
| cool/cold | 0.1562500 | 0.3076923 |
| dark/black | 0.2564103 | 0.1562500 |
| difficult/impossible | 0.5312500 | 0.7948718 |
| fat/obese | 0.1025641 | 0.0937500 |
| funny/hilarious | 0.0937500 | 0.1025641 |
| good/perfect | 0.7179487 | 0.5312500 |
| good/excellent | 0.3125000 | 0.5128205 |
| happy/ecstatic | 0.1025641 | 0.0937500 |
| happy/delighted | 0.0312500 | 0.0512821 |
| hard/unsolvable | 0.6153846 | 0.4375000 |
| hot/boiling | 0.3750000 | 0.2820513 |
| hot/scalding | 0.1282051 | 0.1562500 |
| hungry/starving | 0.2500000 | 0.0769231 |
| large/gigantic | 0.3846154 | 0.0937500 |
| loud/deafening | 0.1875000 | 0.2307692 |
| old/ancient | 0.2307692 | 0.1250000 |
| palatable/delicious | 0.5000000 | 0.4545455 |
| poor/destitute | 0.1538462 | 0.0312500 |
| pretty/beautiful | 0.0312500 | 0.1025641 |
| pretty/gorgeous | 0.1025641 | 0.0625000 |
| quiet/silent | 0.1250000 | 0.3076923 |
| quiet/inaudible | 0.4615385 | 0.4375000 |
| red/scarlet | 0.2812500 | 0.3333333 |
| scared/petrified | 0.1282051 | 0.0625000 |
| sleepy/asleep | 0.6562500 | 0.7692308 |
| small/tiny | 0.2307692 | 0.0937500 |
| smart/brilliant | 0.0000000 | 0.1250000 |
| soft/mushy | 0.4375000 | 0.2051282 |
| special/unique | 0.0512821 | 0.0937500 |
| tasty/delicious | 0.0000000 | 0.0000000 |
| thick/impenetrable | 0.2500000 | 0.3846154 |
| tough/impossible | 0.7435897 | 0.5000000 |
| ugly/hideous | 0.0625000 | 0.0769231 |
| understandable/articulate | 0.1025641 | 0.0312500 |
| unhappy/miserable | 0.0937500 | 0.0769231 |
| warm/hot | 0.6410256 | 0.1562500 |
| wary/scared | 0.1250000 | 0.1794872 |

**Table 3.** Experiment 5 SI rates.

| Scale | Exp5 SI Neutral | Exp5 SI Biased |
|---|---|---|
| attractive/stunning | 0.1833333 | 0.1694915 |
| big/huge | 0.1916667 | 0.0677966 |
| big/enormous | 0.1583333 | 0.1355932 |
| bright/brilliant | 0.0847458 | 0.1129032 |
| calm/meditative | 0.1000000 | 0.0338983 |
| cheap/free | 0.9277778 | 0.9152542 |
| cold/frosty | 0.1833333 | 0.0847458 |
| cold/freezing | 0.3250000 | 0.0593220 |
| cool/cold | 0.3916667 | 0.3220339 |
| dark/black | 0.3250000 | 0.1271186 |
| difficult/impossible | 0.7833333 | 0.7401130 |
| fat/obese | 0.1083333 | 0.0423729 |
| funny/hilarious | 0.0916667 | 0.1101695 |
| good/perfect | 0.7055556 | 0.6779661 |
| good/excellent | 0.3500000 | 0.4830508 |
| happy/ecstatic | 0.2250000 | 0.1101695 |
| happy/delighted | 0.0666667 | 0.0338983 |
| hard/unsolvable | 0.6000000 | 0.5988701 |
| hot/boiling | 0.5555556 | 0.3983051 |
| hot/scalding | 0.2583333 | 0.2881356 |
| hungry/starving | 0.2333333 | 0.1271186 |
| large/gigantic | 0.3416667 | 0.1525424 |
| loud/deafening | 0.3083333 | 0.2203390 |
| old/ancient | 0.3166667 | 0.1186441 |
| palatable/delicious | 0.5081967 | 0.5178571 |
| poor/destitute | 0.2083333 | 0.1016949 |
| pretty/beautiful | 0.0916667 | 0.0762712 |
| pretty/gorgeous | 0.1333333 | 0.1949153 |
| quiet/silent | 0.2166667 | 0.2203390 |
| quiet/inaudible | 0.6222222 | 0.5988701 |
| red/scarlet | 0.3666667 | 0.2203390 |
| scared/petrified | 0.1500000 | 0.1101695 |
| sleepy/asleep | 0.8166667 | 0.7175141 |
| small/tiny | 0.2083333 | 0.1355932 |
| smart/brilliant | 0.0983607 | 0.1428571 |
| soft/mushy | 0.4416667 | 0.2372881 |
| special/unique | 0.0333333 | 0.0423729 |
| tasty/delicious | 0.0338983 | 0.0322581 |
| thick/impenetrable | 0.3666667 | 0.3050847 |
| tough/impossible | 0.8277778 | 0.7740113 |
| ugly/hideous | 0.1000000 | 0.1101695 |
| understandable/articulate | 0.1000000 | 0.0762712 |
| unhappy/miserable | 0.1583333 | 0.0593220 |
| warm/hot | 0.6111111 | 0.2372881 |
| wary/scared | 0.1416667 | 0.1525424 |

**Table 4.** Cohen's d (Experiment 5).

| Scale | Cohen's d Neutral | Cohen's d Biased |
|---|---|---|
| attractive/stunning | 0.4729376 | 0.3763651 |
| big/huge | 0.2897728 | 0.2307059 |
| big/enormous | 0.9046430 | 0.6051680 |
| bright/brilliant | 0.3509475 | 0.4899217 |
| calm/meditative | -0.0609421 | -0.0030722 |
| cheap/free | 0.6582755 | 0.3671337 |
| cold/frosty | 0.1170557 | 0.1131371 |
| cold/freezing | 0.5774044 | 0.1101282 |
| cool/cold | -0.4801559 | -0.0591877 |
| dark/black | 0.7593063 | 0.5141113 |
| difficult/impossible | 0.7438571 | 0.4682022 |
| fat/obese | 0.3393419 | 0.3320644 |
| funny/hilarious | 0.2182463 | 0.4277940 |
| good/perfect | 1.1987234 | 0.5584152 |
| good/excellent | 0.5366780 | 0.7076747 |
| happy/ecstatic | 0.7836990 | 0.7463421 |
| happy/delighted | 0.1011513 | 0.2298266 |
| hard/unsolvable | 0.7589743 | 0.5461049 |
| hot/boiling | 0.9085567 | 0.7862427 |
| hot/scalding | 0.4483090 | 0.5789469 |
| hungry/starving | 0.8786052 | 0.4484844 |
| large/gigantic | 0.6480808 | 0.2376186 |
| loud/deafening | 0.8439826 | 0.6012707 |
| old/ancient | 0.4614332 | 0.4736060 |
| palatable/delicious | -0.1329749 | 0.0477481 |
| poor/destitute | 0.0495170 | 0.0340092 |
| pretty/beautiful | 0.1349479 | 0.2947613 |
| pretty/gorgeous | 0.3868245 | 0.3796296 |
| quiet/silent | 0.1948432 | 0.0718205 |
| quiet/inaudible | 0.1093174 | 0.0494731 |
| red/scarlet | -0.3589848 | -0.0897868 |
| scared/petrified | 0.8287819 | 0.6125419 |
| sleepy/asleep | 0.4931831 | 0.1599763 |
| small/tiny | 0.2381023 | 0.0158797 |
| smart/brilliant | 0.4785733 | 0.7197897 |
| soft/mushy | -0.1313133 | 0.2522070 |
| special/unique | -0.3893946 | 0.2691355 |
| tasty/delicious | 0.2707256 | 0.5005002 |
| thick/impenetrable | 0.4477085 | 0.5014195 |
| tough/impossible | 0.6524549 | 0.5047153 |
| ugly/hideous | 0.5449292 | 0.3413082 |
| understandable/articulate | -0.0270152 | -0.0492938 |
| unhappy/miserable | 0.5015280 | 0.4669700 |
| warm/hot | 0.2230397 | 0.1008658 |
| wary/scared | -0.4474763 | -0.0516081 |

**Table 5.** Boundedness and extremeness.

| Scale | Boundedness | Extreme |
|---|---|---|
| attractive/stunning | Not Bounded | Extreme |
| big/huge | Not Bounded | Extreme |
| big/enormous | Not Bounded | Extreme |
| bright/brilliant | Not Bounded | Extreme |
| calm/meditative | Bounded | Non |
| cheap/free | Bounded | Non |
| cold/frosty | Not Bounded | Extreme |
| cold/freezing | Not Bounded | Extreme |
| cool/cold | Not Bounded | Non |
| dark/black | Bounded | Non |
| difficult/impossible | Bounded | Extreme |
| fat/obese | Not Bounded | Extreme |
| funny/hilarious | Not Bounded | Extreme |
| good/perfect | Bounded | Extreme |
| good/excellent | Not Bounded | Extreme |
| happy/ecstatic | Not Bounded | Extreme |
| happy/delighted | Not Bounded | Extreme |
| hard/unsolvable | Bounded | Extreme |
| hot/boiling | Not Bounded | Extreme |
| hot/scalding | Not Bounded | Extreme |
| hungry/starving | Not Bounded | Extreme |
| large/gigantic | Not Bounded | Extreme |
| loud/deafening | Not Bounded | Extreme |
| old/ancient | Not Bounded | Extreme |
| palatable/delicious | Not Bounded | Extreme |
| poor/destitute | Not Bounded | Extreme |
| pretty/beautiful | Not Bounded | Extreme |
| pretty/gorgeous | Not Bounded | Extreme |
| quiet/silent | Bounded | Non |
| quiet/inaudible | Bounded | Extreme |
| red/scarlet | Bounded | Non |
| scared/petrified | Not Bounded | Extreme |
| sleepy/asleep | Bounded | Non |
| small/tiny | Not Bounded | Extreme |
| smart/brilliant | Not Bounded | Extreme |
| soft/mushy | Bounded | Non |
| special/unique | Bounded | Extreme |
| tasty/delicious | Not Bounded | Extreme |
| thick/impenetrable | Bounded | Extreme |
| tough/impossible | Bounded | Extreme |
| ugly/hideous | Not Bounded | Extreme |
| understandable/articulate | Not Bounded | Non |
| unhappy/miserable | Not Bounded | Extreme |
| warm/hot | Not Bounded | Non |
| wary/scared | Not Bounded | Non |

## Appendix II

In collecting degree estimates for weaker scalar terms (e.g. *The employee is smart*), we needed to ensure that participants did not enrich the adjective with SI (*smart but not brilliant*) and were instead providing a response based on the literal meaning. To do this, we combined the sliding scale task with the inference task. For the analysis, we fitted two mixed effects models (neutral and biased subjects respectively) predicting the sliding scale data from whether SI was calculated as a categorical fixed effect. In both models the predictor was scaled sum coded. Random intercepts by item, as well as by-condition random slopes were included. Results revealed that degree responses were indeed significantly different depending on the inference task response: when a participant had responded "Yes" in the inference task (associated with SI calculation), degree responses were on average lower. This was the case for both neutral subjects (Fig. 6, $\beta = -3.8$, $SE = 0.56$, $t = 6.8$, $p < 0.0001$) and biased subjects (Fig. 7, $\beta = -8.6$, $SE = 0.78$, $t = 11.02$, $p < 0.0001$). The finding that SI calculation corresponds to lower degrees follows from it being an upper-bounding inference; what these results show is that participants judged an employee to be lower on the smartness scale when he was *smart but not brilliant* than when he was *smart*.

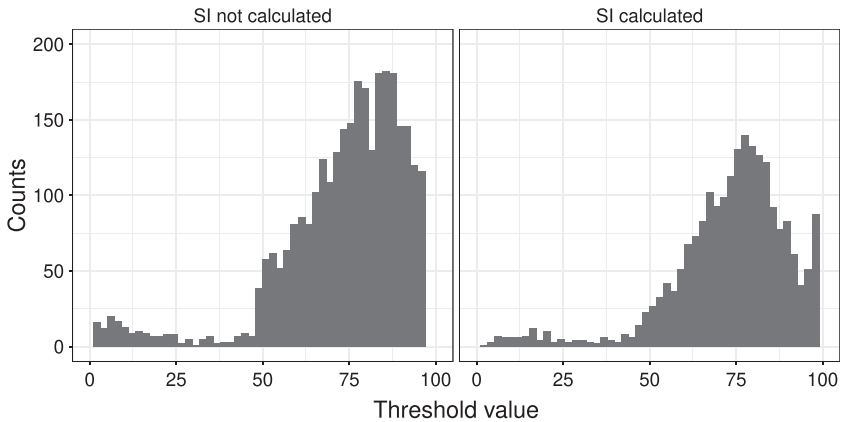


**Figure 6.** Distribution of degree estimates in the weak neutral condition, depending on the result of the following inference task (SI calculated v. not).

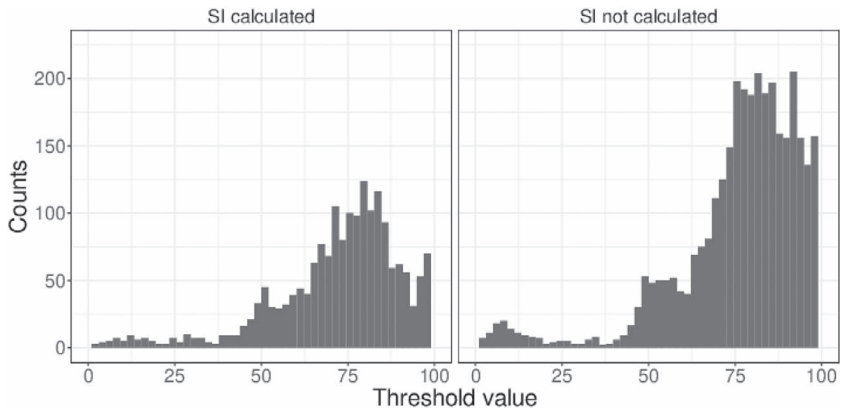


**Figure 7.** Distribution of degree estimates in the weak biased condition, depending on the result of the following inference task (SI calculated v. not)


*Journal of Semantics*, 2025, 1–30
https://doi.org/10.1093/jos/ffaf002
Article