A corpus-based examination of Scalar Diversity

Chao Sun[1,2], Ye Tian[3], Richard Breheny[2]

[1]School of Chinese as a Second Language, Peking University

[2]Department of Linguistics, University College London

[3]Wluper Ltd, London, UK

Author Note

All data analysis code, and research materials are available at https://github.com/sunchaosc/corpus-scalar.git

This research was supported by the LeverhulmeTrust grant RPG-2018-425. Chao Sun was supported by Alexander von Humboldt Fellowship.

Correspondence concerning this article should be addressed to Chao Sun, School of Chinese as a Second Language, Peking University, Yiheyuan Road 5, 100871, Beijing, China. E-mail: chaosun@pku.edu.cn

Abstract

The phenomenon of scalar diversity refers to the well-replicated finding that different scalar expressions give rise to scalar implicatures at different rates. Previous work has shown that part of the scalar diversity effect can be explained by theoretically motivated factors. Although the effect has been established only in controlled lab-based experiments, there has been a tendency to assume that the marked differences in inference rates that have been observed reflect differences to be found in naturally occurring discourse. We explore whether this is the case by sampling actual language usage involving a wide range of scalar expressions. Adopting the approach in Degen (2015), we investigated the scalar diversity effect in a corpus of Twitter data we constructed. We find that the phenomenon of scalar diversity attenuates significantly when measured in a corpus-based paraphrase task. Although the degree of 'scalar diversity' varies, we find that factors derived from theories of scalar implicature can consistently explain nearly two-thirds of the observed variation. As for the remaining variation, we hypothesise that it may be due to a high level of uncertainty about whether adjectival scalar expressions should undergo scalar enrichment.

Keywords: pragmatics, Scalar implicature, Scalar diversity, corpus

A corpus-based examination of Scalar Diversity

A speaker who utters (1) often implies that not all of the students passed the exam. The implicature from 'some' to 'not all' is called a scalar implicature (SI).

(1) Some of the students passed the exam.

(2) All of the students passed the exam.

The SI arises when an alternative like (2), where 'some' is replaced by 'all', is contextually relevant. For example, if (1) is an answer to the question whether all of the students passed, then the fact that the speaker could have but did not utter (2) suggests that it should be excluded. This gives rise to the SI. Conversely, if (1) is an answer to the question if some of the students passed, then (2) is less relevant to the question and the SI is less likely to arise (Fox & Katzir, 2011; Geurts, 2010; Grice, 1967).

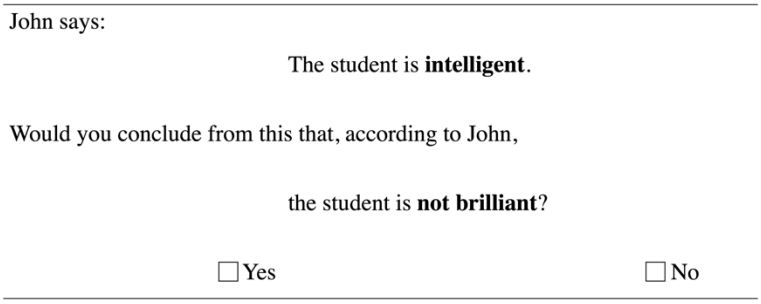
There is a variety of scalar expressions beyond 'some', which include adjectives (e.g., 'warm' often implies 'not hot'), modals ('possible' often implies 'not certain'), and verbs ('try' often implies 'not succeed'). In all of these cases, the implicature is widely assumed to be derived by the same mechanism as the 'not all' implicature, that is, via the exclusion of relevant alternatives. On this basis, an implicit assumption in empirical research on SIs is that different scalar expressions should generate similar outcomes in standard tests.

However, a consistent finding in recent work suggests that some scalar expressions are more likely to give rise to SIs than others (Beltrama & Xiang, 2013; Doran et al., 2009; Van Tiel et al., 2016). For example, van Tiel and colleagues compared the derivation rates of SIs for more than forty scalar expressions using an inference paradigm. In their study, participants read a statement containing a scalar expression and had to decide whether the speaker implied the corresponding SI, which is the negation of a sentence where the scalar expression is replaced by a stronger alternative – see Figure 1. van Tiel and colleagues found that the rates of acceptance varied from 4% to 100%, with quantifiers and modal expressions having higher rates than verbs and adjectives, and with rates among the

adjectives varying widely. This variation in SI rates is often referred to as the Scalar Diversity effect. So far, this effect has been replicated using different tasks, such as the 'Literal Lucy' task in Doran et al. (2009), the consistency test in Simons & Warren (2018), and the inference task with a gradient, *0-100* response (Sun et al., 2018) or with a variety of adjective scales (Gotzner et al., 2018a, 2018b).

Figure 1

Sample item used in van Tiel et al. (2016)'s inference task



The scalar diversity effect raises the question of what factors could contribute to this great variation in the derivation rates of SIs. In particular, it is puzzling that certain scalar expressions, particularly adjectives, yielded surprisingly low rates in van Tiel et al.'s study. It is also puzzling that, in the same inference task, quantifiers and modals yielded high SI rates, relative to their rates in more commonly used verification tasks. We believe that part of the answer comes from the elicitation method in van Tiel et al., where a binary judgment is used in an inference task. For several well-motivated reasons (Franke & Jäger, 2016; Goodman & Frank, 2016), judgments about whether a SI is available may be graded. In an inference task, a participant gives a 'yes' response when they are confident that the speaker intends an enriched understanding where the alternative is excluded. If a participant is presented with a stimulus and is not so confident that the SI is intended, they may respond 'no', even if they can see that a SI could be part of the meaning. In other words, rates of 'no' response do not

solely reflect the proportion of participants who see only the unenriched, literal meaning, but 'no' may be a back-off response, reflecting uncertainty about whether the sentence carries the scalar enrichment or not. Support for this line of thinking comes from the inference-task study presented in Sun et al. (2018), who used the 43 scalar expressions from van Tiel et al. (2016) and two-thirds of the same items. The main difference between the two inference tasks lay in the response type. Sun and colleagues used a 0-100 sliding scale, which allowed participants to express their uncertainty by choosing mid-range values. They found much less variation in rates between the strongly rated quantifiers and modals and the lower rated adjectives and verbs (see Appendix B). Nevertheless, Sun and colleagues still found wide variation in rates among these scalar items and their ranking did correlate with the ranking reported in van Tiel et al. (2016).

The similarities and differences in outcomes of van Tiel et al. and Sun et al. speak to two types of question that are not clearly distinguished in previous discussions of scalar diversity effects: Are scalar inferences derived at different rates in actual everyday language use?; and, to the extent that scalar inferences are derived at different rates, can we explain any variation in terms of factors predicted from theories of scalar implicature? As we will review below, the second question has already been explored in some depth. As to the first question, there has been a tendency to infer from previous results that such great variation reflects real world differences. However, this assumption is yet to be tested more directly. In fact, the method in much previous research has been to present a small number of rather minimal items per scalar expression. Moreover, where a scalar diversity effect has clearly emerged, the target sentences have been presented with minimal or no context, as illustrated in Figure 1 above. While van Tiel et al. present some very good motivations for adopting this more minimal method in designing laboratory tasks, several authors have commented that absence of actual contexts should leave us wary of conclusions we draw about the degree of variation uncovered (McNally, 2017; Ronai & Xiang, 2021). Furthermore,

the small sample of items involved in these studies should lead us to treat the results of statistical analyses with caution (Button et al., 2013; Vadillo et al., 2016). In order to get a better sense of any actual variation in the rates of scalar implicature derivation from these scalar expressions and to put theoretical explorations on better statistical footing, we present here research based on a more extensive sample of naturally occurring items. Before turning to the details of our study, we consider the second question mentioned above, namely how to account for any variation that we find in naturally occurring data.

When considering accounts of the inter-expression variation found in previous inference tasks, we would like to make a distinction between factors which bear directly on the derivation of the scalar implication, factors which bear on the derivation of competing, or conflicting implications and other more methodologically motivated factors. Factors of the first kind are considered in van Tiel et al. (2016), where the authors take into account the explanation of SI as exclusion of alternatives. They consider how the asserted sentence and the sentence containing the alternative are related. They find that a measure of semantic distance between the scalar term and the alternative can explain a small amount of variance, and also whether the underlying scale is 'bounded'. In a related line of research, Sun et al. (2018) consider the extent to which scalar expressions can undergo local scalar strengthening (Bergen et al., 2016), as where the interpretation of 'some' is strengthened to mean *some and not all*. Sun et al. also looked at scale homogeneity, which is a measure of the extent to which a scalar expression or the presented alternative may be understood relative to several scales[1]. For example 'brilliant', which is presented as the alternative to 'intelligent' in van Tiel et al. (2016), can be understood in terms of measures of kindness, skilfulness, etc. in addition to an intelligence scale (McNally, 2017; Hu et al., 2022). This may

[1] Participants were asked to rate the degree of naturalness of the construction 'S but not W', with S being the alternative and W being the scalar expression. For example, 'The student is brilliant but not intelligent'. A lower rating of naturalness indicates a higher degree of scale homogeneity.

have affected which alternative was considered in derivation. Together, these factors have been shown to account for a significant amount of the variance in a regression model of inference task rates, but by no means all of it.

Another kind of factor is one involved in deriving competing or conflicting implicatures. Gotzner et al. (2018a,b) have shown that, particularly for adjectival expressions used in inference tasks, a different kind of enrichment is possible, which is known as *negative strengthening*. Negative strengthening typically occurs when an adjective related to an extreme point on a measure scale is negated. For example, 'The student is not brilliant' can be understood to mean that the student is rather dull. In general, negative strengthening involves negation of one predicate which results in the negation of a weaker predicate on the same measure scale. Gotzner and colleagues argue that inference task probes containing such adjectives as an alternative may be construed in this way, thus contradicting the quoted assertion, 'the student is intelligent'. They created a measure of the ease of negative strengthening for the items in van Tiel et al.'s study and found that this can account for some of the variance in elicited inference task rates. Furthermore, in a modified inference task, they found that using a question probe which blocks negative strengthening (e.g., would you conclude from this that, according to Mary, he is intelligent but not brilliant?) reduced the scalar diversity effect.

Another type of pragmatic enrichment, which has been dubbed as strengthening UP by Sun et al. (2019) (see also Bergen et al., 2016), may also play a role in lowering the rates of SIs in the inference task, especially for adjectives. UP enrichment results from an inference about the contextual standard for the weaker expression on a degree scale. Predicates that are understood relative to a degree scale often have an implicit, context dependent standard. For example, 'tall' can denote different degree intervals on a scale of heights, depending on whether, in the context, the domain of application is buildings, ten-year-old children, etc. To strengthen UP a scale is to set the standard of application higher

than normal. This is often because the speaker provides cues such as special intonation or exclamatives (like 'that is tall, man'), So, for example, in the case of 'intelligent', a not uncommon pragmatic enrichment is to raise the standard of application so that the resulting interpretation is closer to that of 'brilliant'. In such a case, an UP enrichment would result in an enrichment that is not clearly consistent with the probed implication, 'the student is not brilliant'. Hence, where applied, UP enrichment would lower the rates of acceptance for SIs. Sun et al. operationalised UP enrichment and found that scalar adjectives are particularly susceptible to this type of enrichment. Moreover, this factor, like negative strengthening, could account for a significant amount of variance in responses in the inference task.

Methodological factors also contribute to the inter-expression variation found in previous inference tasks. For example, it is difficult to control or take into account the contextual relevance of the SI[2]. In experimental settings, when there is no information about the intended relevance of the utterance, participants tend to infer context based on the linguistic stimuli presented to them (Bergen & Grodner, 2012; Breheny et al., 2006; Tian et al., 2010). Previous inference tasks often present participants with a speaker's statement with little information about its relevance. In such cases, the contextual relevance of the SI is influenced by the type of probe question used in the inference task, as shown by Sun & Breheny (2022). Furthermore, in standard inference tasks, the relevance of the SI is influenced by the likelihood that participants will infer a question about the stronger alternative after reading a statement containing a scalar expression. This factor also contributes to the scalar diversity effect, as shown by Ronai & Xiang (2021).

[2] Pankratz & van Tiel (2021) operationalised the relevance of the SI as the co-occurrence frequency of the scalar constructions such as α *but not* β, where α is the weaker expression and β is a stronger scalemate. They focused on adjective scalar expressions and predicted the SI rates collected from a standard inference task using a number of factors, including their measure of relevance. They found that this measure of relevance was a significant predictor in explaining the variance, but a large amount of variance still remained unexplained.

In this paper, we aim to investigate the scalar diversity effect with a sample of stimuli that better reflect actual experience. To this end, we adopted the corpora and web-based methods used in Degen (2015). In Degen's study, participants' judgments on scalar implicatures were collected from naturally occurring sentences containing 'some'. The stimuli were extracted from the Switchboard corpus of telephone dialogues (Godfrey, Holliman & MaDaniel, 1992). Participants were presented with the critical utterance together with ten utterances from the immediately preceding discourse context. In our study, instead of extracting occurrences of scalar expressions from existing corpora, we constructed a corpus of English Twitter texts containing different scalar expressions. The maximum length of the retrieved tweets is 140 characters. These texts provide sentential context for the scalar expressions investigated in our study. Although perhaps shorter than the ten-utterance context of Degen's study, the sampled tweets are on average longer than the laboratory stimuli of many previous studies. In addition, Twitter texts tend to be designed by their authors to get across a message in the short format. Finally, the sampled set of texts also provides a much more diverse range of items for determining the inference rates of a given scalar expression.

Our sampling of a real-world corpus should, first and foremost, give a better indication of whether, and to what extent, scalar diversity is a robust phenomenon. In addition, our study should go some way to address the factors that lead to competing or conflicting implications by using a paraphrase task. We do not present participants with a sentence that could be negatively strengthened (such as, 'The student is not brilliant'), rather our probe incorporates the negated alternative with the assertion ('The student is intelligent but not brilliant') and so is not open to negative strengthening. In general, context should provide more information about which, if any, strengthening is intended. As a result, UP enrichment is to some extent controlled as well. Finally, as our method allows for more graded judgments about the presence of SIs, we expect that the very wide range of rates

found in van Tiel et al.'s binary judgment task will not be replicated. This expectation is based on comparison with the results in Sun et al. (2018), who used a 0-100 scale, mentioned above. In sum, this corpus-based exploration will give researchers a better sense of the size of the scalar diversity effect in naturally occurring language use, and by addressing competing implications, it will help us explore any observed variance in a better controlled setting.

We tested a subset of the 43 scalar expressions from van Tiel et al., as many as our methodology would allow (explained in footnote 3). In what follows, we will first describe the collection and annotation of the Twitter corpus, then we will present the paraphrase task that measures the rates of SIs of different scalar expressions, and finally we will compare the findings of our study with findings reported by van Tiel et al. (2016) and Sun et al. (2018).

Creating a tweet corpus

The collection of texts

We selected 28 scales from the 43 scales tested in van Tiel et al. (2016). Our selection of scalar expressions includes quantifiers (2 scales), adverbs (1), modals (2) and adjectives (23)[3]. For each of these 28 scales, we extracted texts containing the weak scalar expression from Twitter using Twitter API. In the texts containing scalar adjectives, the

[3] Please see Appendix A for the full list of scales. We did not select auxiliary verbs and main verbs (e.g., <may, will>; <participate, win>) due to implementation issues of the paraphrase paradigm. The comparison sentences of these scales would be constructed differently from the rest. For instance, for a sentence containing 'may' such as "The lawyer may come", its comparison sentence would be "The lawyer may come, but it is not the case that the lawyer will come". In addition, seven adjectives were not selected either due to the infrequent occurrence, e.g., 'unsettling' from the scale <unsettling, horrific>, or due to the infrequent use of the relevant word sense, e.g., in Twitter texts, 'cool' from the scale <cool, cold> is often used to express impressive feeling rather than a fairly low temperature. Based on the SI rates reported in van Tiel et al.'s Exp 2, the variance of the full set (43 scales) was 0.08 and the variance of the subset (28 scales) was 0.09. Levene's test showed that the variances of the two sets were equal ($p = .56$) between the two sets. Thus, we did not inadvertently test a set of scalar expressions with less variability in inference rates.

adjective could be predicative or attributive, as shown in (3). We will discuss later that these different uses of adjectives have an impact on the rates of SIs.

(3) a. The singer is *attractive*.

b. I saw an *attractive* singer.

All texts were retrieved from Twitter users in the United States. The maximum length of the text is 140 characters (the upper limit set by Twitter). To ensure that the text retrieved consists of at least one single sentence, no texts shorter than 30 characters were included.

The selection of texts

The text selection process was conducted in two steps. First, we used an automatic tagging tool to annotate part-of-speech and syntactic information for each tweet text. Based on the tagging results, we applied hand-written rules and regular expressions to identify and exclude texts where scalar implicatures were unlikely to arise due to part of speech or syntactic structure (Noveck et al., 2002). Second, we conducted a word sense disambiguation task and excluded texts labelled as inappropriate or contained scalar expressions in irrelevant senses. Below, we provide the details of the text selection process.

R-assisted automatic tagging and exclusion

Texts were tagged using the GATE Twitter part-of-speech tagger (Derczynski et al., 2013). We first excluded cases where the scalar expression and its alternative were of different parts of speech. Consider, for example, the adjective scale <hard, unsolvable>. We excluded cases where 'hard' was used as an adverb, as in 'work hard'. We then excluded cases in which the scalar expression appeared in environments where the SI is either unavailable or less likely to arise, as shown in Table 1. We also excluded cases that contained certain syntactic constructions known to block SIs. For instance, we adopted Degen (2015)'s criterion that a 'some'-NP headed by a singular count noun (e.g., 'some guy') should be

excluded. Furthermore, we excluded cases where the scalar expression was part of an idiom or phrase (e.g., 'special' in 'special force', 'special edition').

Table 1

Environments that discourage scalar implicature (the scalar expression is in bold).

Environment	Example
under negation	We're not **hungry**.
conditional antecedents	If the weather was **warm**, we would have some people over for a small party in our backyard.
wh-questions	What type of **intelligent** promoter releases the entire amount before the artists arrive at the venue?
polar questions	Do you get **adequate** vitamin D?
questions with auxiliary verbs	Would you add **some** girlfriends to the page?

Word sense disambiguation task

Scalar expressions, especially adjectives, are likely to be polysemous. We consulted the Merriam-Webster dictionary and found that 20 of the expressions investigated here have at least two different but related senses. For example, 'old' from the adjective scale <old, ancient> means "existing for a long time" in (4) and means "previous" in (5). 'Ancient' is a valid alternative to 'old' only in (4). Thus, cases like (5) need to be excluded.

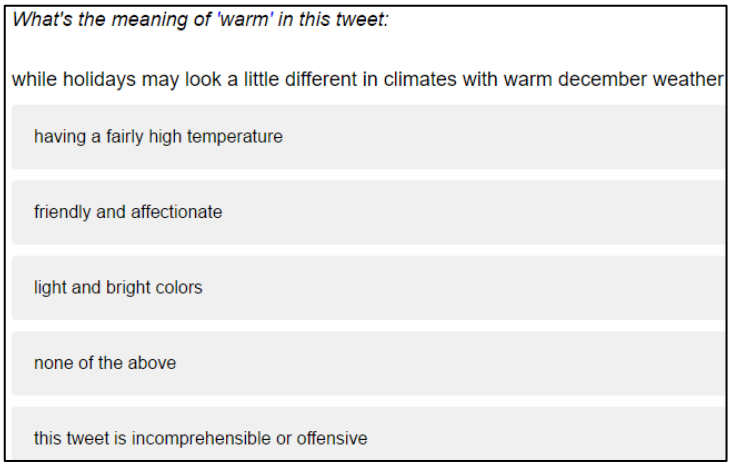
(4) I'm in an **old** abandoned train station w/ a translator working on the script.

(5) That means my **old** boss has been approaching a breakdown for the last 2 years.

To obtain word sense annotations of polysemous scalar expressions in the collected texts, we conducted a word sense disambiguation task on Amazon Mechanical Turk. In this task, US-based workers read a text containing a scalar expression. They then had to choose the meaning of the scalar expression from a set of options. Figure 2 shows an example item.

Figure 2

Word sense disambiguation task example item.



The dataset consisted of 4000 texts, with 200 texts per scale and a total of 20 scales[4]. 80 Mechanical Turk workers were recruited and each annotated 50 texts of one scalar expression. Based on the workers' annotations, we excluded texts that were labelled as incomprehensible or offensive and texts where the scalar expression was used in a sense that invalidates the alternative. After this final exclusion, our corpus consists of 3075 English Twitter texts. We randomly selected 50 texts for each scalar expression and used them as the stimuli for the paraphrase task.

The corpus-based paraphrase task

Methods

Participants

500 participants (mean age: 37, standard deviation: 11, 218 females) were recruited via the Amazon Mechanical Turk and were paid $0.4 for their participation. These participants were located in the United States and had a 95% approval rate on tasks previously performed for other requesters. Participants were asked to indicate their native language, but they were paid regardless of their answer to this question. This experiment

[4] The task did not include 'some', 'few', 'sometimes', 'difficult', 'intelligent', 'memorable', 'scare', 'wary'.

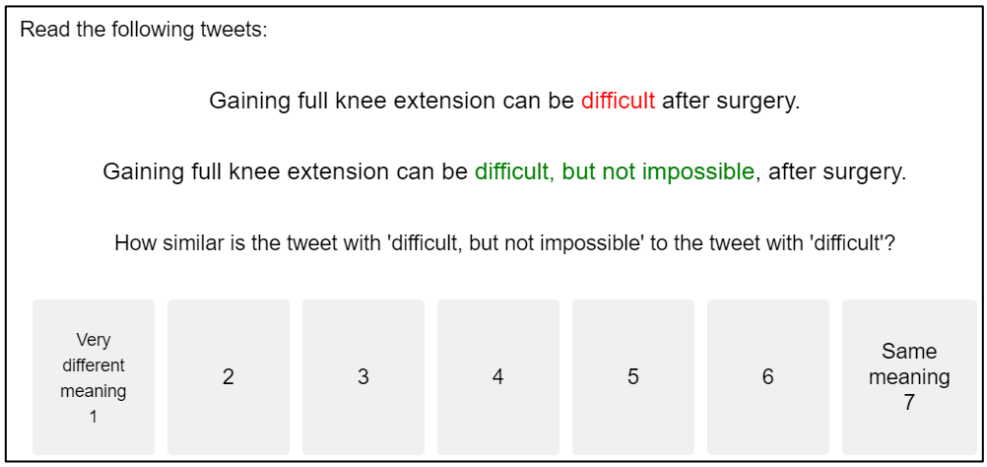
was approved by the local research ethics committee. Participants were provided with an electronic version of informed consent before taking part.

Materials and procedure

The task mirrored Degen (2015)'s study. Figure 3 shows an example item. Participants read a target sentence that contained a scalar expression 'X' (in red), and a comparison sentence, in which the SI 'but not Y' (in green) was inserted after 'X'. Participants had to decide how similar these two sentences are on a 1 (very different meaning) to 7 (same meaning) scale. The higher the similarity rating, the greater the strength of the SI. Two practice trials were included to ensure that the participants understood the instruction and to encourage them to use the full range of the rating scale. The sentences in the practice trials are provided in (6) and (7). We instructed the participants to choose a high value in (6) and a lower value in (7).

Figure 3

Paraphrase task example item



(6) a. And sometimes my German shepherd just growls at my empty bathroom.

b. And sometimes, but not always, my German shepherd just growls at my empty bathroom.

(7) a. Yes, but the fundamental issue is the need to provide adequate funding and joined up thinking.

b. Yes, but the fundamental issue is the need to provide adequate, but not good, funding and joined up thinking.

There were 50 items per scale, and 1400 items in total. Each participant judged 28 items, one item per scale. Each item received 7 to 11 judgments (average 10). This variation was due to the randomisation of assigning participants to items using the Qualtrics survey platform.

Transparency and openness

We report all data exclusions and all measures in the study. All data analysis code, and research materials are available at https://github.com/sunchaosc/corpus-scalar.git. Data were analysed using R (R Core Team, 2020) and the package *ggplot2* (Wickham, 2016). The design and analysis of this study were not pre-registered.

Results

Eight participants were removed for having a native language other than English. 492 participants were thus included in the following analyses. To enable comparisons with previous studies, we rescaled the SI ratings in the corpus-based paraphrase task and Sun et al.'s inference task to the range of [0, 1] [5], which is consistent with the range used in van Tiel et al.'s inference task.

Comparing rankings

We first compared the rankings of scalar expressions in the paraphrase task with those reported in van Tiel et al. (2016) and Sun et al. (2018). In each study, we ranked 28 scalar expressions based on their implicature rates/ratings in a descending order. Kendall's τ

[5] The original ratings were transformed to a scale of 0-1 using the formula y= $\left(\frac{x-x_{min}}{x_{range}}\right)n$, where x is the observed rating, x_{min} is the minimal observed rating, x_{range} is the difference between the maximum potential rating and the minimum potential rating on the rating scale, and n is the upper limit of the rescale rating.

revealed significant correlations between these rankings (τ = 0.63, p < .001, with van Tiel et al.'s binary task; τ = 0.50, p < .001, with Sun et al.'s continuous task). Furthermore, Kendall's W showed significant agreement on the rankings among all three studies (w = .86, p < .001). These results show that there is a great similarity in the ranking of the scalar expressions across studies with different response types.

Comparing variances

Figure 4 shows the variability of by-scale mean SI rates in three studies. While the three studies have similar overall mean rates, the by-scale means[6] were more packed in the paraphrase task than in the van Tiel et al.'s and Sun et al.'s studies. We conducted Levene's test to compare the equality of variances between the three studies. There was a significant difference in the variances between the current study and the previous ones (F(1,54) = 25.49, p < .001, with van Tiel et al.'s task; F(1,54) = 14.91, p < .001, with Sun et al.'s task). Table 2 describes the variation in each study in terms of standard deviation, range and variance.

The difference in the variability observed in our current study and in Sun et al.'s inference task using continuous measurement suggests that the attenuated SI effect is not solely a consequence of switching from binary to continuous measurement. Rather, it is likely that the decreased SI effect is due to accounting for competing enrichments, such as negative strengthening and UP enrichment, and incorporating additional contextual information[7].

[6] The mean SI rate for each scale is shown in the bar chart in the Appendix B.

[7] Gotzner et al. (2018) conducted an inference task that was modified to block negative strengthening. Participants read 'He is intelligent' and had to respond yes/no to the question 'would you conclude from this that, according to Mary, he is intelligent but not brilliant?'. The study included 70 adjective scales, 23 of which were also tested in our paraphrase task. We ran Levene's test to compare the results of the two studies and found that the variance was significantly reduced in the corpus-based paraphrase task (F(1,44) = 4.84, p = .03). Given that Gotzner et al.'s task asked for a binary response, we cannot explain the reduced variance as solely due to the inclusion of more contextual information in our task. However, it is likely that using corpus-based context further helps to reduce variability.

Figure 4

The diamond represents the overall mean for each study, along with bootstrapped 95% confidence intervals. Each dot corresponds to a by-scale mean value.

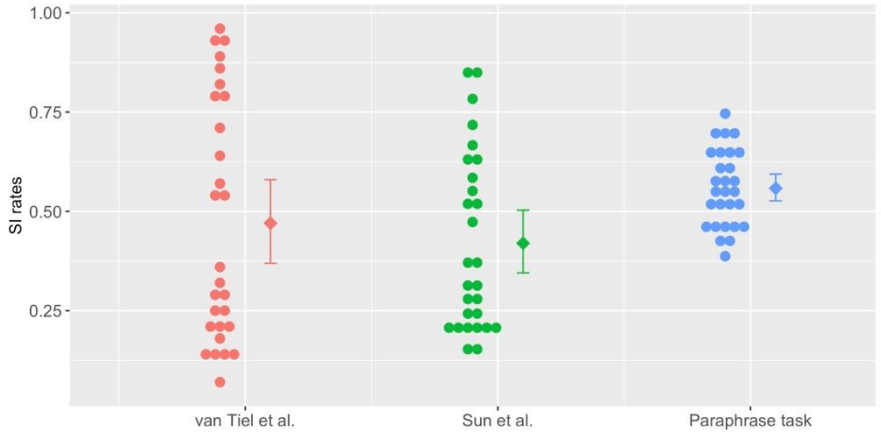


Table 2

Mean and measures of variance for each study.

	van Tiel et al.	Sun et al.	Paraphrase task
Mean	0.47	0.42	0.56
Standard deviation	0.30	0.22	0.10
Range	0.07-0.96	0.14-0.86	0.39-0.75
Variance	0.09	0.05	0.01

To further compare the variability across scales, Figure 5 shows the distribution of SI rates for the non-adjective and adjective scales tested in all three studies. There were 5 non-adjective scales[8] and 23 adjective scales. We can see that the difference in variability across studies are primarily driven by the adjectival scales. For non-adjective scales, the overall mean SI rate was lower in the paraphrase task, yet the variability was similar to that found in previous studies. Whereas for adjectival scales, not only were the overall mean rates higher

[8] Non-adjective scales include some, few, sometimes, allowed, possible.

in the paraphrase task, the by-scale mean rates also varied in a mild manner compared to what was found in previous studies.

Figure 6 shows the distribution of by-item means for each adjective scale in the current study. It shows that most of the adjective scales have item means clustered around the midpoint, rather than a mixture of high and low means.

Figure 5

The distribution of by-scale SI rates for each study. The left panel shows the distribution in non-adjective scales, and the right panel shows the distribution in adjective scales. Each dot corresponds to a scale. The diamond represents the overall mean with bootstrapped 95% confidence intervals.

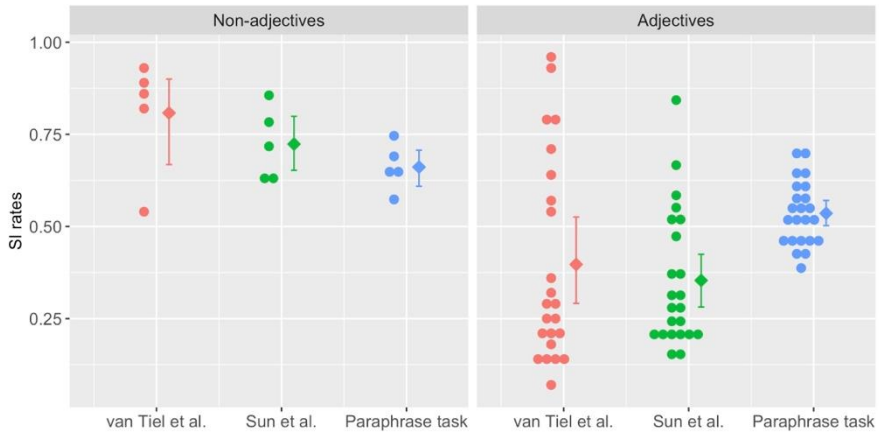
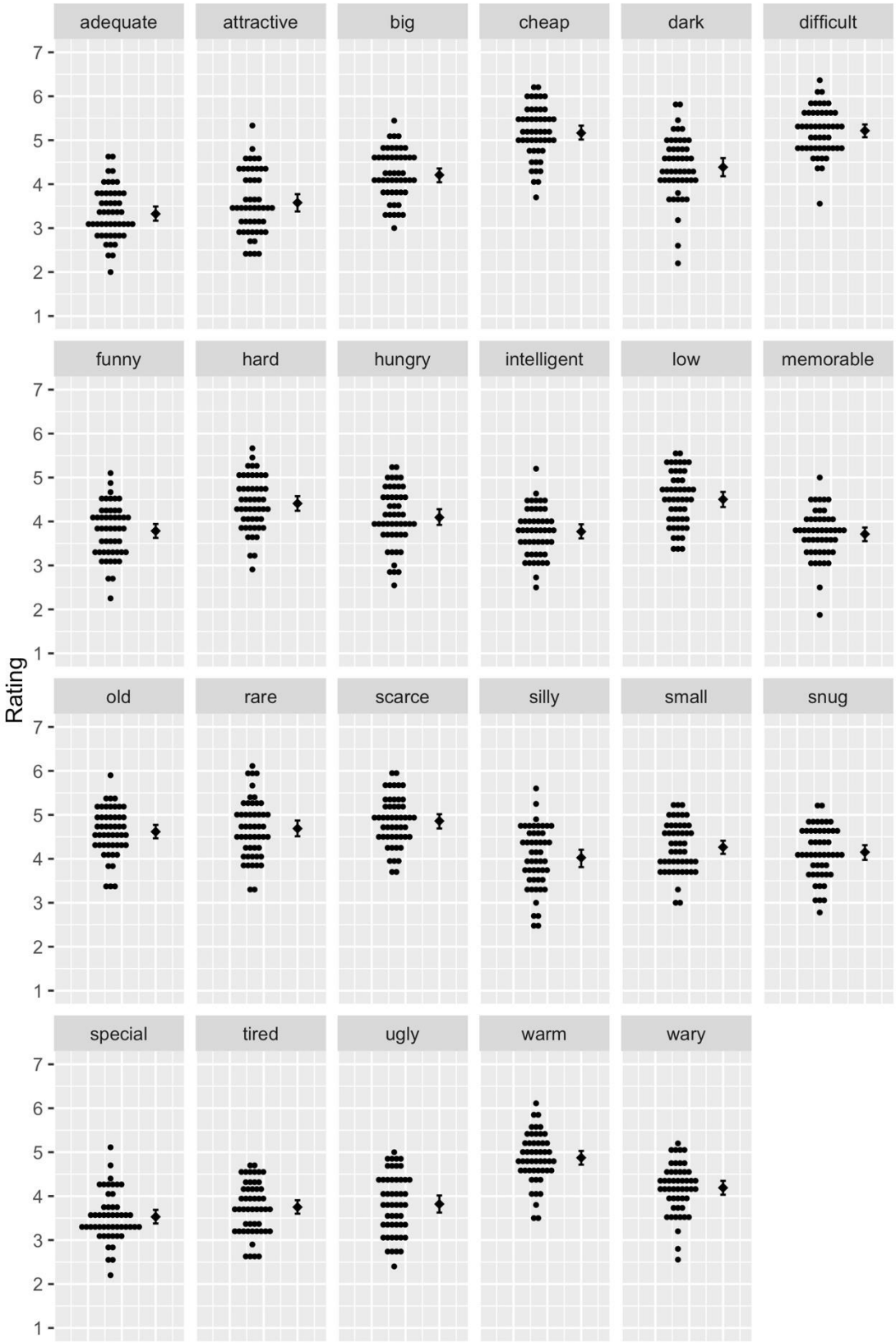


Figure 6

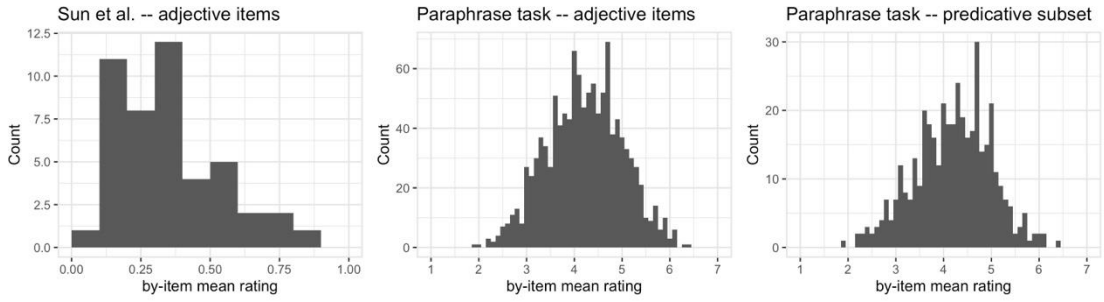
The distribution of by-item means for each adjective scale. The diamond represents the overall mean with bootstrapped 95% confidence intervals.


adequate
attractive
big
cheap
dark
difficult
funny
hard
hungry
intelligent
low
memorable
old
rare
scarce
silly
small
snug
special
tired
ugly
warm
wary
Rating
1
2
3
4
5
6
7

We now turn to a closer comparison of the SI rates of the adjective scales in Sun et al. and the current study. In both studies, data were collected from a continuous or semi-continuous rating scale. Figure 7 shows the distributions of by-item mean ratings for the adjectival scales tested in both studies. In Sun et al. (2018), the distribution for adjectival scales is right skewed (skewness: 0.86), with data clustering around the lower end of the rating scale. By contrast, the distributions in the current study, particularly the distribution for predicative adjectives, are left skewed (all adjectives:−0.05, predicative: −0.12). Thus, participants' ratings have moved from a tendency to a without-SI interpretation in a laboratory task without context, to a tendency to rate the corpus-based items more in the middle of the scale, albeit with a centre just above the median. To the extent that the corpus language sample reflects users' experience of scalar adjectives, the results suggest that scalar adjectives are oftentimes found to be ambiguous between the with-SI and without-SI interpretations.

Figure 7

Distributions of by-item mean ratings for adjective scales in Sun et al. (2018) (left column) and the current study (middle and right columns).



Note. The middle panel shows the distribution of ratings for all adjective items tested in the paraphrase task, the right panel shows the distribution of ratings for items in which the adjective was used predicatively.

Regression analyses

We fitted a multiple linear regression to test whether the factors that directly affect the derivation of SIs could explain the variability observed in the corpus-based paraphrase task. The dependent variable was the mean by-scale SI rating, and the predictors were the factors investigated in van Tiel et al. (2016) and Sun et al. (2018). This model accounted for 67% of the variance ($R^2 = 0.77$, $F(8,19) = 7.96$, $p < .001$). We then extracted the results of the same 28 scales from both van Tiel et al. (2016) and Sun et al. (2018), and fitted the same linear regression model to each subset. We found that the regression model explained a similar proportion of the total variation across all studies. The regression results are summarised Table 3.

The regression analyses indicate that boundedness and enrichability remain significant predictors across studies. This suggests that the relationship between SI rates and these factors is consistent across studies, even though different paradigms were used to measure the implicature rates. As for semantic distance and homogeneity, only one factor remains significant, and which factor it is varies between datasets. Sun et al.'s measure of homogeneity was highly correlated with van Tiel et al.'s measure of semantic distance, suggesting that these two measures may be colinear, with only one contributing significantly to the model. Given the differences in items, sample size and measurements, the effect of homogeneity on the SI rates may be stronger than semantic distance in one study, while the opposite may be true in another study. Finally, in contrast to previous studies, grammatical class becomes a significant predictor. A possible explanation is that in previous studies, grammatical class captured the partial effect of homogeneity in addition to its own effect. In the paraphrase task, homogeneity became a significant predictor. This increased explanatory power of homogeneity allows us to examine the effect of grammatical class independently of scale homogeneity. In other words, grammatical class is a significant predictor because it accurately reflects its true effect in the model of the paraphrase task.

Table 3

Results of multiple linear regression analyses.

	Paraphrase task			van Tiel et al.			Sun et al.		
Predictors	*β*	*SE*	*p*	*β*	*SE*	*p*	*β*	*SE*	*p*
(Intercept)	0.27	0.15	0.10	-0.88	0.44	0.06	-0.34	0.37	0.37
Association strength	-0.00	0.00	0.16	0.00	0.00	0.60	0.00	0.00	0.48
Grammatical class	0.13	0.06	**0.03**	0.32	0.16	0.06	0.06	0.14	0.67
Word frequencies	-0.02	0.01	0.18	-0.03	0.04	0.39	-0.04	0.03	0.18
Semantic relatedness	0.10	0.06	0.15	0.02	0.18	0.90	-0.06	0.15	0.72
Semantic distance	0.02	0.02	0.19	0.14	0.05	**0.02**	0.09	0.04	0.07
Boundedness	0.07	0.02	**0.01**	0.38	0.07	**<0.01**	0.23	0.05	**<0.01**
Enrichability	0.05	0.02	**<0.01**	0.10	0.05	**0.03**	0.10	0.04	**0.02**
Homogeneity	-0.05	0.02	**0.01**	-0.06	0.06	0.29	-0.04	0.05	0.40
Observations	28			28			28		
R^2 / R^2 adjusted	0.770 / 0.673			0.809 / 0.729			0.753 / 0.649		

Discussion

Research into scalar diversity was initially motivated by an interest in the question whether all lexical scales are equal in the extent to which they give rise to scalar inference. Taking some earlier studies at face value, the answer seemed to be a resounding no. However, for most previous research the focus has been on what factors contribute to variation, rather than being controlled explorations of the extent of actual variation in everyday language use. We have set out to place scalar diversity research on a firmer methodological footing by undertaking just such an exploration. We presented participants with stimuli which featured different lexical scalar terms, drawn from large samples of naturally occurring texts. In addition, our probe stimuli took into account factors that may

have affected the outcomes in previous studies. First, instead of asking for a binary judgment, we used a 7-point Likert scale to collect participants' judgments about how strongly the use of the weaker term was felt to convey the negation of the stronger term. This type of response makes it possible to capture participants' uncertainty about the interpretation of a scalar expression in context. Second, we adopted the paraphrase paradigm from Degen (2015). This paradigm blocks the possibility that participants' responses are influenced by negative strengthening. Lastly, our corpus stimuli provided more diverse and richer contexts in which the scalar expressions would be interpreted. In general, more contextual information potentially make clear which inference is intended in each item. This is particularly important for adjectival scales, as adjective scalars are susceptible to UP enrichment and the problem of scale homogeneity.

Once we controlled for these factors, we found that the effect of scalar diversity was much smaller compared to what had been reported in previous work. Specifically, comparing the ranges of average ratings in van Tiel et al.'s inference task results and the current study's, we have moved from ratings that cover almost the whole range (0.07-0.96), to a much less variable set of ratings (0.39-0.75).

The difference in variation between the current study and van Tiel et al.'s inference task further demonstrates that the diversity effect attenuates when participants were asked for graded judgments. This suggests that our judgments about whether a SI is conveyed are rarely absolute and are mostly gradient. This finding has general methodological implications for how to interpret participants' binary judgments in an inference task. While a 'yes' response clearly indicates that participants consider the SI as part of the intended meaning, a 'no' response is more complicated. It could indicate that the SI is considered not part of the intended meaning, or it could be used as a back-off response when participants are uncertain about the intended meaning. Thus, in comparison to binary responses, a

continuous or semi-continuous rating scale provides a better reflection of uncertainty in participants' interpretation.

As to why SI rates are lower for adjectives than for quantifiers and modals, even in real-world contexts, one hypothesis is that there is often not enough information to determine which, if any, kind of enrichment is intended. This hypothesis is based on the rating distributions of the adjective scales, where ratings are clustered around the centre. We assume that since pragmatic inferences are derived with some degree of uncertainty, there exists a minimum threshold of certainty below which Scalar Inferences (or other enrichments) are not considered. Our observations suggest that, for quantifiers and modals, confidence in such inferences often exceeds this threshold and requires less contextual information. For scalar adjectives, however, confidence in such inferences often falls short of the threshold and requires more information. This hypothesis is not directly tested in our work, and it needs further exploration.

The percentage of total variance explained by theoretically motivated factors was consistent across various studies (63% in Sun et al., 66% in Gotzner et al., 67% in our current study). These results suggest that while these factors have a stable effect on the derivation rates, the variance explained by these factors is limited. The unexplained variation may include systematic variance associated with unknown variables and random variance due to measurement error. Like van Tiel et al., we find it hard to identify additional factors that may explain the remaining variance. It is possible that certain situation-specific factors may account for some of the variance, such as the relevance of SIs in the context. However, defining contextual relevance of SIs in a corpus-based study is challenging.

The variability in SI rates decreased considerably from van Tiel et al.'s study to Sun et al.'s and finally to the current study. Our study contributes to the existing literature by investigating the scalar diversity effect in real-world language use and demonstrating that

factors predicted from theories of scalar implicature can explain nearly two-thirds of the variation, regardless of the degree of 'scalar diversity'.

References

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

Bergen, L., & Grodner, D. J. (2012). Speaker knowledge influences the comprehension of pragmatic inferences. *Journal of Experimental Psychology: Learning Memory and Cognition*, *38*(5), 1450–1460. https://doi.org/10.1037/a0027850

Bergen, L., Levy, R., & Goodman, N. D. (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, *9*. https://doi.org/10.3765/sp.9.20

Breheny, R., Katsos, N., & Williams, J. (2006). Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition*, *100*(3), 434–463. https://doi.org/10.1016/j.cognition.2005.07.003

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience*, *14*(5), 365-376.

Degen, J. (2015). Investigating the distribution of some (but not all) implicatures using corpora and web-based methods. *Semantics and Pragmatics*, *8*(0), 11-1–55. https://doi.org/10.3765/sp.8.11

Derczynski, L., Ritter, A., Clarke, S., & Bontcheva., K. (2013). Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data. *Proceedings of the International Conference on Recent Advances in Natural Language Processing, ACL*.

Doran, R., Baker, R. E., McNabb, Y., Larson, M., & Ward, G. (2009). On the Non-Unified

Nature of Scalar Implicature: An Empirical Investigation. *International Review of Pragmatics*, *1*, 211–248. https://doi.org/10.1163/187730909X12538045489854

Fox, D., & Katzir, R. (2011). On the characterization of alternatives. *Natural Language Semantics*, *19*(1), 87–107. https://doi.org/10.1007/s11050-010-9065-3

Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift Fur Sprachwissenschaft*, *35*(1), 3–44. https://doi.org/10.1515/zfs-2016-0002

Geurts, B. (2010). *Quantity implicatures*. Cambridge: Cambridge University Press. https://doi.org/10.1016/j.neuroimage.2009.02.016

Goodman, N. D., & Frank, M. C. (2016). Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. https://doi.org/10.1016/j.tics.2016.08.005

Gotzner, N., Solt, S., & Benz, A. (2018a). Adjectival scales and three types of implicature. *Semantics and Linguistic Theory*, *28*(01), 409. https://doi.org/10.3765/salt.v28i0.4445

Gotzner, N., Solt, S., & Benz, A. (2018b). Scalar diversity, negative strengthening, and adjectival semantics. *Frontiers in Psychology*, *9*(SEP), 191–203. https://doi.org/10.3389/fpsyg.2018.01659

Grice, H. P. (1967). Logic and conversation. william james lectures, harvard university. *Studies in Syntax*, *3*.

Noveck, I. A., Chierchia, G., Chevaux, F., Guelminger, R., & Sylvestre, E. (2002). Linguistic-pragmatic factors in interpreting disjunctions. *Thinking & Reasoning*, *8*(4), 297-326. DOI: 10.1080/13546780244000079

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

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

Simons, M., & Warren, T. (2018). A closer look at strengthened readings of scalars. *Quarterly Journal of Experimental Psychology*, *71*(1), 272–279. https://doi.org/10.1080/17470218.2017.1314516

Sun, C., Benz, A., Gotzner, N., Richard, & Breheny, R. (2019). *Approaching scalar diversity through RSA with Lexical Uncertainty*. [Poster Abstract]. The 8th Experimental Pragmatics conference, University of Edinburgh, UK. https://www.xprag2019.ppls.ed.ac.uk/abstracts/sun2.pdf

Sun, C., Tian, Y., & Breheny, R. (2018). A link between local enrichment and scalar diversity. *Frontiers in Psychology*, *9*(OCT), 2092. https://doi.org/10.3389/fpsyg.2018.02092

Sun, C., & Breheny, R. (2022). The role of Alternatives in the interpretation of scalars and numbers: Insights Insights from the inference task. *Semantics and Pragmatics*, *15*.

Tian, Y., Breheny, R., & Ferguson, H. J. (2010). Why we simulate negated information: A dynamic pragmatic account. *Quarterly Journal of Experimental Psychology*, *63*(12), 2305–2312. https://doi.org/10.1080/17470218.2010.525712

Vadillo, M.A., Konstantinidis, E. & Shanks, D.R. (2016). Underpowered samples, false negatives, and unconscious learning. *Psychon Bull Rev* 23, 87–102. https://doi.org/10.3758/s13423-015-0892-6

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

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Appendix A: Scales used in the corpus-based paraphrase task

<adequate, good>; <allowed, obligatory>; <attractive, stunning>; <big, enormous>; <cheap, free>; <dark, black>; <difficult, impossible>; <few, none>; <funny, hilarious>; <hard, unsolvable>; <hungry, starving>; <intelligent, brilliant>; <low, depleted>; <memorable, unforgettable>; <old, ancient>; <possible, certain>; <rare, extinct>; <scarce, unavailable>; <silly, ridiculous>; <small, tiny>; <snug, tight>; <some, all>; <sometimes, always>; <special, unique>; <tired, exhausted>; <ugly, hideous>; <warm, hot>; <wary, scared>

Appendix B: Mean SI rates of each scale

Mean SI rates of each scale (sorted by van Tiel et al.'s results). *Note.* Ratings from the paraphrase task and Sun et al. were rescaled to the range 0-1.