Language and Cognition (2024), 1–26
doi:10.1017/langcog.2024.36

ARTICLE

Scales and inferences

Shirly Orr[1,2], Mira Ariel[1] and Einat Shetreet[1,2]

[1]Department of Linguistics, Tel Aviv University, Tel Aviv, Israel and [2]Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel
Corresponding author: Shirly Orr; Email: shirlym3@mail.tau.ac.il

(Received 25 June 2023; Revised 13 June 2024; Accepted 19 June 2024)

Abstract
Scalar inferences (SIs) are upper-bounding inferences associated with the use of semantically lower-bounded scalar expressions. One of the current debates regarding these inferences concerns their inferential pattern, specifically whether SIs are uniform or diverse across scales. This study follows the work on scalar diversity yet introduces two changes: First, we reexamine, from a different perspective, two structural properties of scales identified as accounting for SI diversity (boundedness and distance). Second, we analyze our data using both traditional regression analysis and complementary cluster analysis. The regression analysis demonstrates that our reexamination of the structural properties provides a more effective model, which also emphasizes the relationship between boundedness and distance. Specifically, we propose that boundedness fixes distance. The cluster analysis demonstrates two scale types: given-scales, which have an entrenched scalar construal, trigger SIs robustly; and volatile-scales, which have a fluctuant scalar construal, trigger SIs inconsistently. Building on these two scale types, we propose a necessary distinction between the conceptualization of a scale, which is *diverse* across different scales, and the actual derivation of the SI, which is *uniform* for all scales, once a scale has been construed. This distinction, we propose, explains how diversity can coexist alongside uniformity.

Keywords: boundedness; distance; scales; scalar diversity; scalar inferences

1. Introduction

Scalar inferences (SIs) are pragmatically derived upper-bounding inferences associated with the use of semantically lower-bounded scalar expressions, such as 'some' or 'warm' (Geurts, 2010; Hirschberg, 1985; Horn, 1972, 1989). This definition encompasses two components: scales and inferences. In this study, we provide further evidence regarding two structural properties of scales previously identified as relevant to SI derivation, namely boundedness and distance (van Tiel et al., 2016), to improve our understanding of (i) the conceptualization of scales on the one hand, and (ii) the inferential mechanism associated with their use on the other.

Scales are sequences of expressions ordered according to increasing informativity (Geurts, 2010, p. 50). Informativity, a rather vague notion, is sometimes discussed in terms of strength (mostly in adjectival scales such as <*beautiful, stunning*>), and on other occasions in terms of entailment relations (for logical scales such as <*some, all*>) (see Hirschberg, 1985, p. 50; Horn, 1989, p. 233 regarding why relying on entailment relations alone may hinder an account of quantity inferences associated with non-logical expressions). Practically, for expressions to be ordered in terms of informativity, they must share the same underlying property (e.g., 'attractiveness' for 'beautiful' and 'stunning'), while expressing different degrees of that property (Kennedy & McNally, 2005; Paradis, 2001; see also the requirement of monotonicity, Horn, 1989). Obviously, if two expressions denote the same interval of the same underlying property, they do not constitute a scale and, in fact, should be understood as synonyms. That is, a 'scalar construal'—the conceptualization of two scalar expressions as plotted on the same scale as well as having some distance between them—is needed for expressions to constitute a scale. Given that scalar expressions are analyzed as semantically lower bound, distance is measured between the two expressions' lowerboundary (also see Gotzner et al., 2018, p. 10).

These attributes—a shared underlying property on the one hand and distance on the other—have led researchers to discuss scalar expressions as weak/strong scale mates. Thus, two scalar expressions are considered weak/strong scale mates when the lower bound of the weak scalar expression denotes a lower degree of the underlying property than the lower bound of the other, stronger scalar expression (and conversely for scales with negative monotonicity). For example, in the scale <*beautiful*, *stunning*>, the lower bound of 'beautiful', the weaker scalar expression, denotes a lower degree of the underlying property than the lower bound of 'stunning', the stronger scalar expression. Conversely, in the negative monotonic scale <*cool*, *cold*>, 'cool', the weaker scalar expression, denotes a higher lower bound of the underlying property than 'cold', the stronger scalar expression.

In light of the analysis of scalar expressions as semantically lower-bounding only, they are said to cover the semantic meaning of their stronger scale mate. That is, because 'beautiful' is assumed to have only a lower bound, the stronger scale mate 'stunning' is a possible interpretation of 'beautiful' ('beautiful and possibly stunning'). The inference associated with the use of weaker scalar expressions, the SI, is the rejection of these stronger interpretations. That is, the SI adds a pragmatic upper-bounding interpretation to the semantic lower-bounded meaning of the relevant expression. Consider (1).

(1) Sue: "Jill is beautiful."

According to the standard theory, using 'beautiful', the weaker term on the <*beautiful, stunning*> scale, tends to trigger a rejection of the stronger alternative, 'stunning'. In such cases, 'beautiful' expresses 'beautiful but not stunning' (for a different view, see Ariel, 2004, 2015; Devlesschouwer, 2019; Koenig, 1991; Orr et al., 2023). The rationale behind this inference is based on the maxim of Quantity (Grice, 1975, 1989; Horn, 1972, 1989) and rests on the following reasoning (see Geurts, 2010; Soames, 1982): The addressee assumes that the speaker could have expressed the stronger statement (i.e., 'Jill is stunning' for example (1)) but did not do so because they do not believe that the stronger statement holds. This inference

https://doi.org/10.1017/langcog.2024.36 Published online by Cambridge University Press

is known as a weak implicature, $\neg BEL_s(p)$, where 'p' stands for the stronger alternative. If the addressee also assumes that the speaker is competent regarding the stronger alternative (i.e., the speaker believes *p* ('Jill is stunning') or $\neg p$ ('Jill is not stunning') in example (1), which translates to $BEL_s(p) \vee BEL_s(\neg p)$), a stronger inference follows (see Van Rooij & Schulz, 2004). The stronger inference is the one associated with the SI, namely, the speaker believes that the stronger alternative does not hold (i.e., that Jill is *not* stunning for example (1)). This inference is also known as a strong implicature, $BEL_s(\neg p)$.

Ever since Doran et al. (2009, 2012), and even more so following van Tiel et al. (2016), the uniformity of SIs has been questioned. By extending the variety of the scales tested, van Tiel et al. demonstrated that the presumed uniformity of SIs, primarily based on the investigation of a limited number of scales (most notably, <*some, all*>, <*or, and*>), cannot be generalized to all scales. Specifically, they found that weaker scalar expressions trigger the rejection of their stronger scale mates to varying degrees. This finding challenged the implicit uniformity assumption, namely that SIs are derived uniformly across different scales. Van Tiel et al. further proposed that the structural properties of the scale can predict some of this diversity. Relevant structural properties were boundedness (i.e., whether the stronger scale mate denotes an endpoint) and distance (between the lower bounds of the two scale mates). The results of this seminal study were replicated in many other works that explored several other factors, such as polarity and extremeness (Benz et al., 2018; Gotzner et al., 2018; Simons & Warren, 2018; Sun et al., 2018, inter alia).

In the current study, we follow the work of van Tiel et al. (2016) on scalar diversity but offer a new perspective on both the methodological measures of boundedness and distance as well as the question of uniformity versus scalar diversity. Experiment 1-a is a replication of van Tiel et al.'s Experiment 2. All in all, we were able to replicate the SI pattern found in van Tiel et al. (2016) (Section 2.1). Experiment 1-b addresses the effect of negative strengthening on this SI pattern; we conclude that this is unlikely to underlie scalar diversity (Section 2.2). We then proceed to measure two of the structural properties shown to predict scalar diversity: boundedness and distance (van Tiel et al., 2016). Experiment 2 is an empirical investigation of boundedness. Contrary to the common view, we measure boundedness as a bias rather than an absolute either-or property (Section 2.3). Experiment 3 focuses on distance. Our experiment builds on van Tiel et al.'s Experiment 4 but modifies the task so that it measures distance without presupposing scalarity, which we argue was the case in the original study and replications of it (Section 2.4). In Section 3, we analyze the data using traditional regression analysis and a complementary cluster analysis, which, as yet, has not been used to study scalar diversity. Based on an interaction shown by the regression analysis, we propose that boundedness fixes distance and hence imposes a scalar structure (Section 4.2.1). Furthermore, based on the cluster analysis, we propose a distinction between two scale types: given scales, with an entrenched scalar structure, show high SI rates, and volatile-scales, with a fluctuant structure, exhibit low SI rates (Section 4.2.2). We argue that this distinction allows us to account for diversity across scales (depending on whether a scale has actually been construed), while adhering to the uniformity of the derivational process (once the scale has been construed) (Section 4.2.3).

https://doi.org/10.1017/langcog.2024.36 Published online by Cambridge University Press

2. Experiments

2.1. Experiment 1-a: A modified replication of van Tiel et al. (2016)

Experiment 1-a is a modified replication of van Tiel et al.'s (2016) Experiment 2. In their study, van Tiel et al. tested whether previous experimental work, which typically focused on a few scales (most notably <*some, all*>, and <*or, and*>), could be extended to other, not previously tested scales. To that end, they conducted two experiments. In these experiments, participants were presented with a statement containing a weak scalar expression (the target sentence). They were then asked to decide whether they could conclude from that statement that, according to the speaker, the stronger scale mate does not hold. A positive response reflected the derivation of the SI (i.e., the rejection of the stronger scale mate). In van Tiel et al.'s Experiment 1, statements were kept as neutral as possible by using pronouns and generic predicates (e.g., 'She is intelligent'). In their Experiment 2, van Tiel et al. used the same items, but targets included full noun phrases and more specific predicates (e.g., 'The assistant is intelligent'; see Figure 1).

We replicated van Tiel et al.'s Experiment 2 but slightly modified the materials in two ways. The first modification involved rewording the target sentences in two scales: <*participate*, *win*> and <*start*, *finish*>. We suspected that the context of the original statements only supported a weak inference (i.e., the speaker does not believe that the stronger scalar expression holds, $\neg BEL_s(p)$). To promote the strong inference (i.e., the speaker believes that the stronger scalar expression does not hold, $BEL_s(\neg p)$), we used meaningful descriptions at the subject/agent position for both scales as well as information about the nature of the competition to the target sentences of the <*participate*, *win*> scale (e.g., 'My cousin participated in the Olympic games' compared to 'The runner participated' in van Tiel et al. (2016); for the complete list see the Supplementary Materials). The second modification was adding another item for the <*possible*, *certain*> scale. The additional item aimed to avoid 'Yes' responses prompted by common knowledge rather than the actual derivation of an SI. In the former case, this would mean, for example, that a selection of 'not certain' for 'success is possible' will result from the widespread belief that success is never certain rather than from the SI derivation (i.e., the derivation of 'not certain' when 'possible' is expressed). Henceforth, we mark the original scale in van Tiel et al.'s study as <*possible*, *certain*>$_{-a}$, and ours as <*possible*, *certain*>$_{-b}$.

2.1.1. Method

Participants. Thirty participants were recruited using Prolific (Palan & Schitter, 2018). All participants were native US English speakers with no history of cognitive

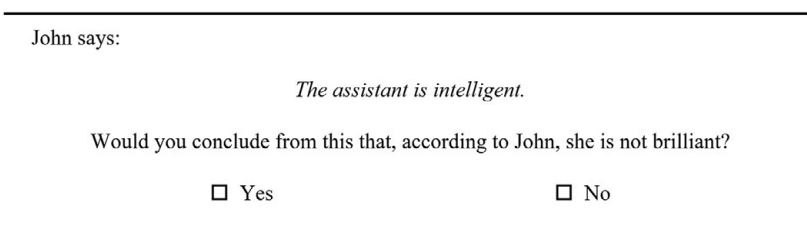

Figure 1. A trial example from van Tiel et al.'s (2016) Experiment 2 using the <*intelligent, brilliant*> scale. A 'yes' response indicates that an SI was drawn.

https://doi.org/10.1017/langcog.2024.36 Published online by Cambridge University Press

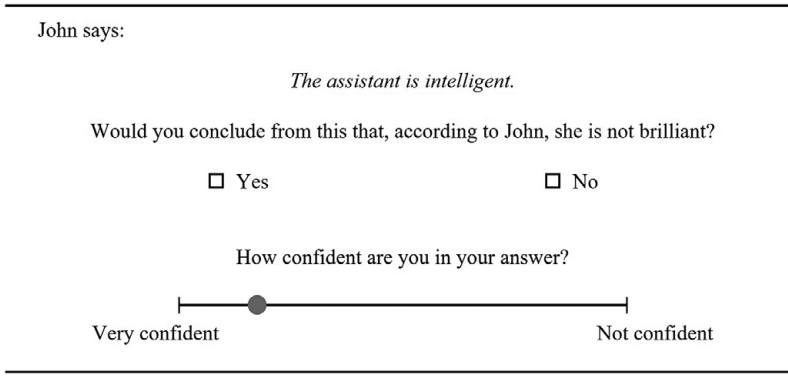

Figure 2. A trial example from Experiment 1-a using the <*intelligent, brilliant*> scale. A 'yes' response indicates that an SI was drawn.

impairment. All participants responded correctly to the control items; therefore, all were included in the analysis (age range: 18–45, $M = 32.5$, $SD = 7.53$, 15 females). All participants gave informed consent prior to participation.

Materials and procedure. We used the same experimental items and controls as van Tiel et al.'s (2016) Experiment 2: quantifiers (2), adverbs (1), auxiliary verbs (2), main verbs (6), adjectives (32, +1 additional <*possible, certain*>$_{\text{-b}}$ scale]), and controls (7), with modifications as detailed in the introduction. For each experimental scale, we used the same three statements as van Tiel et al.'s (2016) Experiment 2. Thus, for example, the scale <*intelligent*, *brilliant*> could have appeared in any of the following statements: 'The assistant is intelligent', 'That professor is intelligent', or 'This student is intelligent'. Control items (from van Tiel et al.) were either antonym expressions (e.g., <*clean, dirty*>) or unrelated adjectives (e.g., <*tall, single*>). The complete list of statements can be found in the Supplementary Materials.

The procedure largely followed that used by van Tiel et al. with the addition of a continuous confidence question, using a slider between 'very confident' and 'not confident'. We added this question because 'Yes' responses to the main task, which indicate an SI derivation, may be associated with different degrees of confidence (see the use of a continuous scale for SI derivation in Sun et al., 2018). Thus, the measuring of confidence could provide an alternative (indirect) measurement of SI diversity (following a paradigm used in several cognitive studies, e.g., Kanai et al., 2010 for vision and Sternau et al., 2015; and Orr et al., 2017 for language).

Participants saw all scales but only one statement per scale. This setup yielded three lists, with ten participants assigned to each list, in accordance with van Tiel et al. (2016). Items in each list were randomized for each participant, apart from the additional <*possible*, *certain*>$_{\text{-b}}$ scale, which was always the last item in the list (Figure 2).

2.1.2. Results

In this analysis, and throughout the following analyses, we used R version 4.1.3 (R Core Team, 2021), RStudio (RStudio Team, 2020), and the tidyverse package (Wickham et al., 2019). SI rates in Experiment 1-a successfully replicated the

https://doi.org/10.1017/langcog.2024.36 Published online by Cambridge University Press

SI rates in van Tiel et al.'s Experiment 2 ($r = 0.88$, $p < 0.001$). To test the correlation between the studies when using the *exact* same stimuli, we removed the scales for which we made modifications (<*participate*, *win*> and <*start*, *finish*>) and performed a second correlation. As expected, this resulted in a higher correlation rate ($r = 0.94$, $p < 0.001$). All three lists correlated with one another ($rs > 0.86$, $ps < 0.001$), indicating that the different statements either contributed or did not contribute to SI rates in a fairly similar manner. Turning to the three modifications we made to the original design: (1) SI rates in the modified scales increased dramatically compared to van Tiel et al.'s Experiment 2 (from 18% to 46% ($\chi^2(1) = 4.9$, $p < 0.05$) for <*participate*, *win*>, and from 21% to 86% ($\chi^2(1) = 24.2$, $p < 0.001$) for <*start*, *finish*>) (see further discussion in the Supplementary Materials). (2) The new <*possible, certain*>$_{\text{-b}}$ scale did not yield any significant difference from the original <*possible, certain* >$_{\text{-a}}$ in this task. (3) Confidence scores were high for all scales ($M = 78$, $SD = 6.61$, and always above 65). Therefore, we could not derive new insights concerning SI diversity from this measurement. Figure 3 and Table 1 provide a detailed description of these results.

2.2. Experiment 1-b: Eliminating the possible interference of negative strengthening

One of the criticisms leveled against SI diversity concerns the potential interference of negative strengthening. This concern was raised and tested by Benz et al. (2018) with respect to the materials in van Tiel et al. (2016). According to Benz et al., it is possible that while an SI was derived during the target stage, it was obscured by negative strengthening during the task stage. To illustrate, a participant encountering the statement 'He is intelligent' may have derived the SI that Mary believes that he is not brilliant (which should lead to a 'Yes' response). However, when asked, 'Would you conclude from this that, according to Mary, he is not brilliant?' during the task stage, 'not brilliant' was negatively strengthened to 'not intelligent', leading to the selection of the 'No' response. To test this hypothesis, Benz et al. collected the rates of negative strengthening for the stronger scale mates tested in van Tiel et al. using a dedicated task (Figure 4).

Benz et al.'s results, which negatively correlated with SI rates from van Tiel et al. (2016), led them to suggest that the absence of SIs in some scales can, at least in part, be explained by negative strengthening. For example, in van Tiel et al. (2016), the scale <*happy, content*> received an SI rate of 4%. In Benz et al. (2018), this scale received a negative strengthening score of 92%. This example illustrates a (potential) pattern whereby '[…] participants are less likely to endorse a scalar implicature if they apply negative strengthening to the stronger scale-mate' (Benz et al., 2018, p. 195).

It is important to note that even before Benz et al.'s experiment, van Tiel et al. (2016, p. 149) considered the possible effect of negative strengthening on SI rates. However, they rejected this possibility for several reasons. We note two. First, diversity in SI rates is also present in tasks that do not involve the negation of the stronger scale mate (Doran et al., 2009, 2012). Second, negative strengthening seems unlikely because it would go against the explicit target sentence stating that the weaker alternative is the case. To address this concern, we nevertheless performed a test with a task that is similar to Experiment 1-a but cannot

https://doi.org/10.1017/langcog.2024.36 Published online by Cambridge University Press

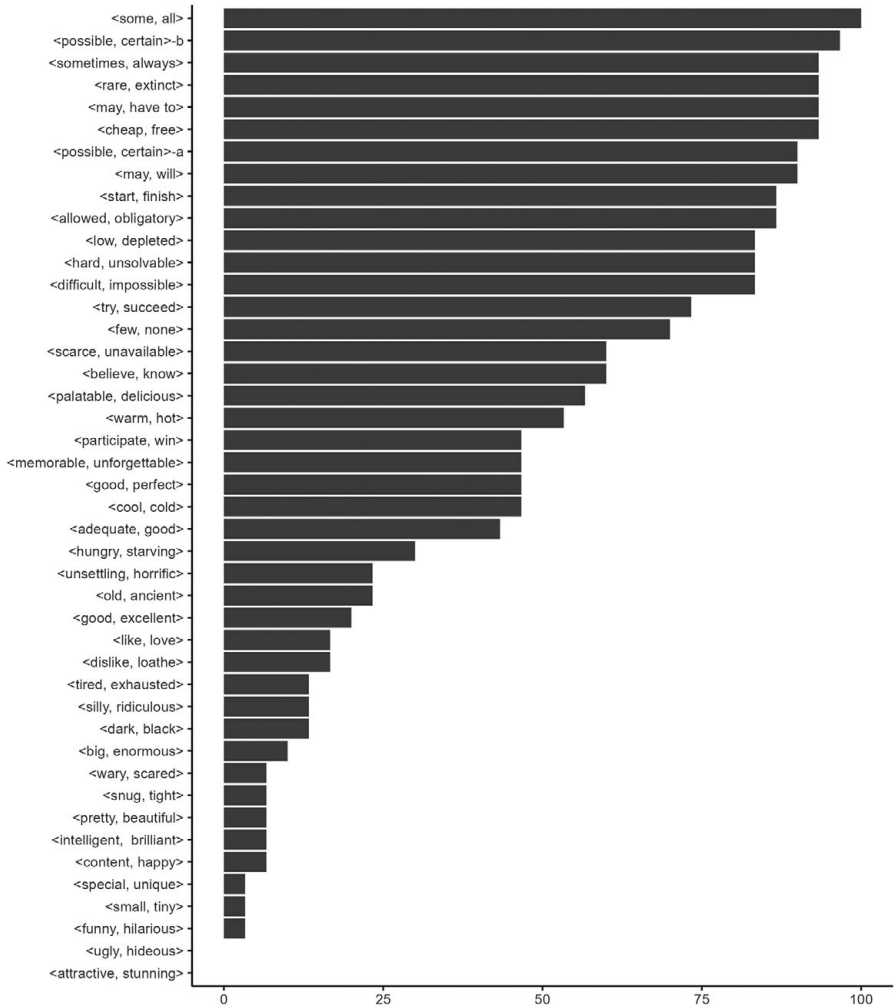

Figure 3. The percentages of 'Yes' responses from our Experiment 1-a (indicating the probability of deriving an SI).

trigger negative strengthening. Our goal was to compare the results of this test with those of Experiment 1-a to understand the contribution, if any, of negative strengthening.

2.2.1. Method

Participants. A different group of 32 participants was recruited using Prolific. All participants were native US English speakers with no history of cognitive impairment. Two participants were excluded from the analysis because they failed in more than two control items. Therefore, 30 participants were included in the final analysis

https://doi.org/10.1017/langcog.2024.36 Published online by Cambridge University Press

Table 1. Results from Experiments 1-a, 2, and 3, as well as the cluster analysis, ordered by SI rates

Scale	SI rate	Boundedness score	Distance score	Cluster
<*some*, *all*>	100	83.98	74.45	1
<*possible*, *certain*>$_{-b}$	96.67	63.13	73.6	1
<*sometimes*, *always*>	93.33	54.42	68.94	1
<*may*, *have to*>	93.33	50.4	79.63	1
<*cheap*, *free*>	93.33	92.74	67.7	1
<*rare*, *extinct*>	93.33	92.6	78.12	1
<*may*, *will*>	90	86.74	64.2	1
<*possible*, *certain*>$_{-a}$	90	54.02	68.44	1
<*start*, *finish*>	86.67	83.39	95.61	1
<*allowed*, *obligatory*>	86.67	66.52	75.86	1
<*difficult*, *impossible*>	83.33	70.14	51.06	1
<*hard*, *unsolvable*>	83.33	82.22	52.84	1
<*low*, *depleted*>	83.33	72.78	35.97	1
<*try*, *succeed*>	73.33	44.14	81.7	1
<*few*, *none*>	70	92.97	88.97	1
<*believe*, *know*>	60	56.51	59.11	1
<*scarce*, *unavailable*>	60	79.1	48.36	1
<*palatable*, *delicious*>	56.67	14.07	64.3	2
<*warm*, *hot*>	53.33	6.42	42.3	2
<*participate*, *win*>	46.67	83.27	75.8	1
<*good*, *perfect*>	46.67	62.61	46.87	1
<*memorable*, *unforgettable*>	46.67	44.47	28.28	2
<*cool*, *cold*>	46.6	4.32	27.97	2
<*adequate*, *good*>	43.33	9.89	48.17	2
<*hungry*, *starving*>	30	34.24	33.3	2
<*old*, *ancient*>	23.33	23.22	23.83	2
<*unsettling*, *horrific*>	23.33	9.56	33.99	2
<*good*, *excellent*>	20	12.45	36.3	2
<*dislike*, *loathe*>	16.67	27.72	16.25	2
<*like*, *love*>	16.67	21.95	34.33	2
<*dark*, *black*>	13.33	46.86	38.18	2
<*silly*, *ridiculous*>	13.33	14.51	26.48	2
<*tired*, *exhausted*>	13.33	7.09	18.07	2
<*big*, *enormous*>	10	21.74	17.77	2
<*content*, *happy*>	6.67	8.01	31.4	2
<*intelligent*, *brilliant*>	6.67	11.32	21.35	2
<*pretty*, *beautiful*>	6.67	12.7	17.54	2
<*snug*, *tight*>	6.67	9.98	10.11	2
<*wary*, *scared*>	6.67	9.22	25.72	2
<*funny*, *hilarious*>	3.33	9.73	13.9	2
<*small*, *tiny*>	3.33	14.3	15.34	2
<*special*, *unique*>	3.33	38.14	23.91	2
<*attractive*, *stunning*>	0	11.28	17.3	2
<*ugly*, *hideous*>	0	8.8	13.31	2

Note: The **SI rate** column provides the results from Experiment 1-a and shows the average likelihood of deriving a scalar inference for each scale; The **Boundedness score** column provides the results from Experiment 2 and shows the mean boundedness score for each scale; The **Distance score** column provides the results from Experiment 3 and shows the mean distance score between scale mates for each scale; The **Cluster** column provides the associated cluster of each scale, determined based on the cluster analysis.

(age range: 18–45, M = 31.5, SD = 7.35, 15 females). All participants gave informed consent prior to participation.

Materials and procedure. We used the same materials and procedure as in Experiment 1-a; the only change was in the task. Here, instead of asking: 'Would you conclude from this that, according to Mary, he is not brilliant?' as in our Experiment 1-a and the

https://doi.org/10.1017/langcog.2024.36 Published online by Cambridge University Press

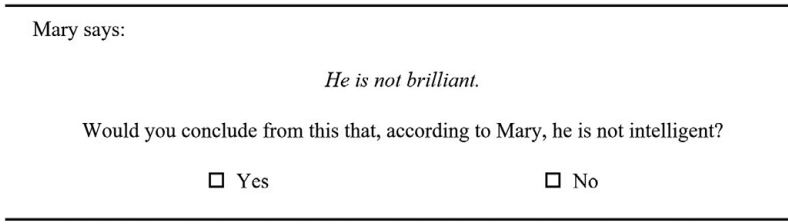

Figure 4. A trial example from Benz et al. (2018) using the <*intelligent, brilliant*> scale.

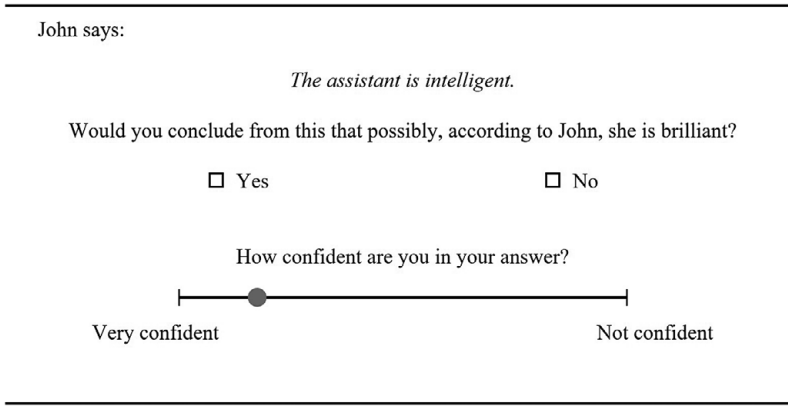

Figure 5. A trial example from Experiment 1-b using the <*intelligent, brilliant*> scale.

original van Tiel et al. (2016) experiment, we asked, 'Would you conclude from this that possibly, according to Mary, he is brilliant?' Now, a negative response, 'No', reflects the derivation of an SI, whereas a positive response, 'Yes', testifies that an SI was not derived. Crucially, by eliminating the 'not' and adding 'possibly' to the task, we were able to eliminate the possible interference of negative strengthening (Figure 5).

We hypothesized that if negative strengthening does not interfere with the derivation of SI rates, Experiment 1-a and this experiment would be (negatively) correlated (because, in this task, the selection of 'No', rather than 'Yes', indicates that an SI is derived). Note that this experiment does not directly test for negative strengthening but only whether SI rates in our Experiment 1-a (and by implication, those in van Tiel et al.) were affected by the presence of 'not', the possible source of negative strengthening.

2.2.2. Results

Our results show that the SI rates of Experiment 1-a and this experiment demonstrate a negative correlation ($r = –0.89$, $p < 0.001$). In other words, SIs were derived with a relatively similar frequency. We thus conclude that it is even less likely that negative strengthening interfered with the SI rates obtained in Experiment 1-a.[1] Therefore, we

[1]In the following sections, we show that some scale mates, such as <*snug, tight*>, were perceived as interchangeable, i.e., as very close in meaning. If so, this offers an alternative interpretation for Benz et al.'s (2018) findings. For example, rejecting 'not tight' when 'snug' was used (the task in van Tiel et al., 2016) is

https://doi.org/10.1017/langcog.2024.36 Published online by Cambridge University Press

will use the results obtained in the previous experiment in the following analyses, wherein we explore two of the structural properties that were found to predict SI derivation, namely Boundedness and Distance.

2.3. Experiment 2: Testing the perceived boundedness of scalar expressions

The first structural property of scales that we explore is boundedness. A bounded scale is a scale in which the stronger scalar expression denotes an endpoint (e.g., 'all' in <*some, all*>, 'free' in <*cheap, free*>). This endpoint indicates that there is nothing, no further alternatives, beyond that point for that scale. Conversely, in non-bounded scales, the stronger scalar expression, similar to the weaker scalar expression, denotes an interval. Whether the scale has (or does not have) an endpoint is relevant to various phenomena in language (see Kennedy & McNally, 2005; Paradis, 2001). One of these phenomena is the derivation of SIs, as found in van Tiel et al. (2016).

To explore the effect of boundedness on the derivation of SIs, van Tiel et al. (2016) divided their scales based on whether the stronger scale mate denotes an endpoint or an interval. They found that the boundedness of the stronger scale mate is the strongest predictor of SIs, accounting for 10% of the variance observed in SI rates. This result was corroborated by Sun et al. (2018), who found that when SI rates were measured on a continuous scale, boundedness accounted for as much as 31% of the variance. These studies, and studies regarding boundedness in general, however, have mostly stipulated which lexical items are bounded (or not) (but see Pankratz & Van Tiel, 2021). We surmised that such an intuitive characterization of boundedness might be subject to errors. For example, the scalar expression 'know' is defined as bounded by van Tiel et al. but as continuous by Paradis (2001, p. 47); also see Kennedy and McNally (2005, p. 365).

Moreover, contrary to the common assumption that boundedness is a binary property, we hypothesized that bounded expressions might exhibit varying degrees of resistance to their interpretation as an interval. That is, we considered (and therefore measured) boundedness as a flexible trait. Our hypothesis was motivated by the availability of 'gradable readings' of what are considered strictly bounded expressions. For example, 'very', which is a scalar modifier, can be added to adjectives that are considered bounded (e.g., 'very certain', 'very true') (see examples and discussion in Paradis (2001) and Kennedy and McNally (2005)). Such uses question the practice of treating boundedness as an absolute either-or property and highlight why exploring it as a bias could be a promising endeavor. In this respect, our investigation also differs from the empirical examination of boundedness performed by Pankratz and Van Tiel (2021), who ultimately divided their items in a binary fashion. The goal of Experiment 2 was, therefore, to quantify boundedness empirically, thereby providing a more objective measure of this structural property.

2.3.1. Method

Participants. A different group of 34 participants was recruited using Prolific. All participants were native US English speakers with no history of cognitive

expected because they are perceived as denoting a close meaning. Similarly, accepting 'not snug' when 'not tight' was used (the task in Benz et al., 2018) is expected for the same reason.

https://doi.org/10.1017/langcog.2024.36 Published online by Cambridge University Press

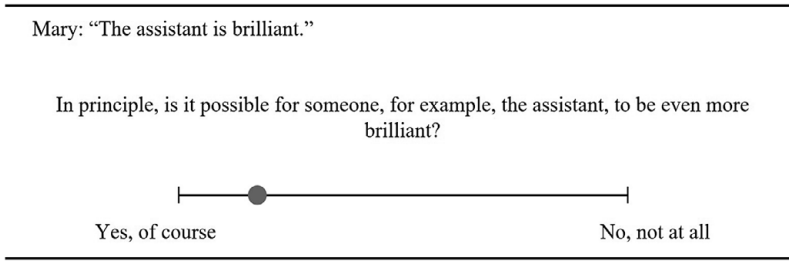

Figure 6. A trial example from Experiment 2 using the <*intelligent, brilliant*> scale. A response towards the left end, 'Yes, of course', indicates that the scalar expression in the target sentence is biased towards a non-bounded conceptualization, and vice versa.

impairment. Four participants were excluded from the analysis because they failed in more than two control items. Therefore, 30 participants were included in the final analysis (age range: 18–45, M = 28.9, SD = 7.65, 15 females). All participants gave informed consent prior to participation.

Materials and procedure. We used the same 44 experimental and 7 control items from Experiment 1-a. For each experimental item, we used the same three statements as in Experiment 1-a. However, we replaced the weak scalar expression with its stronger alternative in these statements (the target sentences) because boundedness is the property of the stronger scalar expression. To measure boundedness, we built on the observation that expressions biased towards boundedness are overall more infelicitous in comparative constructions (see Paradis, 2001, Section 3). We therefore asked participants to indicate, for example, how likely it is for anyone described as 'brilliant' to be even 'more brilliant' than 'brilliant' (Figure 6). Note that while this observation refers to adjectives, we applied it to all items, ensuring that the task was appropriate for non-adjectival scalar expressions.[2]

2.3.2. Results

Boundedness scores in Experiment 2 showed substantial variability (M = 41, SD = 30.66, ranging from the lowest boundedness score of 4.3 for <*cool*, *cold*> to the highest 92.8 for <*few*, *none*>). All three lists correlated with one another (rs > 0.92, ps < 0.001), indicating that the different statements either contributed or did not contribute to boundedness scores in a fairly similar manner. We note here that all items considered bounded by van Tiel et al. received scores higher than the median score (and vice versa; see the ordered boundedness scores in the Supplementary Materials). In fact, a strong point-biserial correlation was observed between our boundedness scores and the either-or distinction assumed in van Tiel et al. (r = 0.9, p < 0.001) (using the ltm package; Rizopoulos, 2006). If we indeed treat boundedness as a categorical property, an interesting pattern is evident. For expressions with boundedness scores higher than the median, the mean was relatively far from the edge of the maximal score (M = 68.23, SD = 17.7). In contrast, for expressions with boundedness scores lower than the median, the mean was relatively close to the

[2]For example, for 'love', a non-adjectival expression biased towards non-boundedness, we asked, 'In principle, is it possible for someone, for example, the manager, to love spaghetti even more?'

https://doi.org/10.1017/langcog.2024.36 Published online by Cambridge University Press

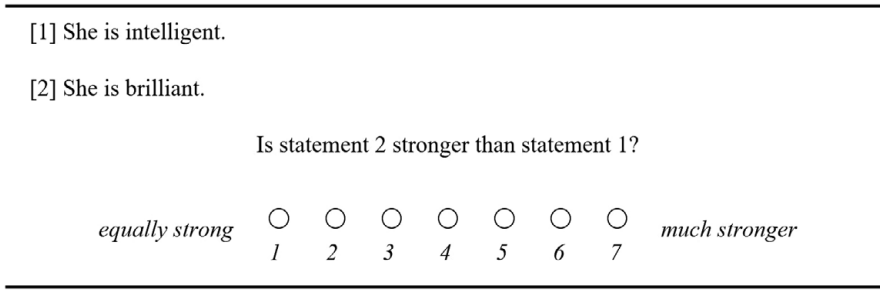

Figure 7. A trial example from van Tiel et al. (2016) Experiment 4, which tested distance using the <*intelligent, brilliant*> scale.

edge of the minimal score (M = 13.75, SD = 7.5). We will develop this point further in Section 4.

2.4. Experiment 3: Testing the perceived distance between scale mates

Following the findings of van Tiel et al. (2016), the second structural property of scales we explore is the perceived distance between two scale mates. To measure this distance, van Tiel et al. asked participants to indicate the extent to which a statement with a stronger scalar expression is stronger than a statement with a weaker scalar expression (Figure 7).

Results from that study showed that distance is a significant predictor of SI rates, albeit negligible, as it accounted for only 3% of the variation in SI rates. We posited, however, that van Tiel et al.'s task made the scalar difference between the two expressions highly pronounced, which may have biased the participants to assume a difference in strength between them. Put differently, it could very well be that by asking how much stronger something is than something else, strength, which was the operationalization of distance, was presupposed. This raises the concern that for *some of the scales* the observed differences in distance were merely biased by the task. As the distance between (the lower bounds of) two scalar expressions is a necessary requirement for a scalar construal, we aimed to measure distance using a task that does not bias for strength/distance. Based on van Tiel et al.'s (2016) Experiment 4, we constructed a similar task but changed it to obtain a measurement of distance that does not presuppose distance and hence scalarity. Specifically, rather than asking about a difference in strength, as van Tiel et al. did, we asked about the interchangeability of the two statements. That is, we were interested in whether, or to what extent, these scale mates indeed denote different intervals on the same scale. We hypothesized that the more two statements are perceived as interchangeable, that is, as denoting the same interval on the same scale, the closer they are in terms of meaning, and vice versa. The goal of Experiment 3 was, therefore, to explore distance without presupposing it.

2.4.1. Method

Participants. A different group of 33 participants was recruited using Prolific. All participants were native US English speakers with no history of cognitive impairment. Three participants were excluded from the analysis because they failed in more

https://doi.org/10.1017/langcog.2024.36 Published online by Cambridge University Press

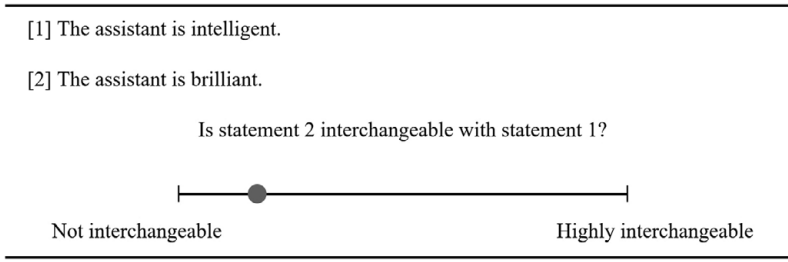

Figure 8. A trial example from Experiment 3 using the <*intelligent, brilliant*> scale. A response towards the left-end, 'Not interchangeable', indicates that the two scale mates are perceived as distant, and vice versa.

than two control items. Therefore, 30 participants were included in the final analysis (age range: 18–45, $M = 30.63$, $SD = 7.12$, 15 females). All participants gave informed consent prior to participation.

Materials and procedure. We used the same 44 experimental and 7 control items from Experiment 1-a. While we used the paradigm from van Tiel et al.'s (2016) Experiment 4, we applied three changes: (1) Whereas van Tiel et al. (2016) used the neutral version of the materials when measuring distance (with the pronouns and generic predicates; Figure 7), we used the contextually richer version of the materials. This was motivated by a desire to make the comparison with our Experiment 1-a more reliable; (2) We used a continuous scale rather than a 1–7 Likert scale (see Dillon & Wagers, 2021 for a discussion); (3) Most importantly, to avoid the presupposition of distance, and hence scalarity, we changed the task used in van Tiel et al. (2016) from 'Is statement 2 **stronger than** statement 1?' to 'Is statement 2 **interchangeable with** statement 1?' (Figure 8).

2.4.2. Results

Scores in Experiment 3 showed substantial variability ($M = 44.7$, $SD = 24.5$; ranging from the highest distance score of 95.6 for <*start, finish*> to the lowest 10.1 for <*snug, tight*>). We note that the scores, which we refer to as 'Distance scores', are reported as the inverse of the interchangeability scores (Distance score = 100 – interchangeability score). Thus, for example, a 4.4 interchangeability score for <*start, finish*> reflects the distance score of 95.6. These scores correlate with the measurement of distance in van Tiel et al. ($r = 0.4$, $p < 0.05$), but note that in Section 4, we provide a more nuanced analysis of this correlation. All three lists correlated with one another ($rs > 0.92$, $ps < 0.001$), indicating that the different statements either contributed or did not contribute to distance scores in a fairly similar manner.

3. General results

3.1. Regression analysis

In the subsequent analyses, we use SI rates taken from Experiment 1-a because the results of Experiment 1-b removed the potential interference of negative strengthening. The first analysis follows that of van Tiel et al. (2016). We performed a binomial mixed effects analysis using the lme4 package (Bates et al., 2015). The SI rates obtained from Experiment 1-a were modeled as a function of boundedness (standardized scores; Experiment 2), distance (standardized scores; Experiment 3),

https://doi.org/10.1017/langcog.2024.36 Published online by Cambridge University Press

Table 2. SI rates modeled as a function of the standardized boundedness and distance scores and the interaction between them in a mixed-effect model

	SI rates		
	Odds Ratios	CI	*p*-value
(Intercept)	1.06	0.78–1.44	0.701
Boundedness	2.09	1.64–2.67	<0.001
Distance	4.11	3.11–5.45	<0.001
Boundedness × Distance	0.55	0.45–0.69	<0.001
Marginal R^2/Conditional R^2	0.49/0.55		

and the interaction between them as fixed factors, as well as the maximal theoretically relevant random effects structure, which included by-participants intercepts and slopes (Barr et al., 2013). To ensure the data met the assumption of collinearity, we applied the Variance Inflation Factors (VIF) function using the CAR package (Fox & Weisberg, 2019). The test showed that multicollinearity was not a concern (Boundedness, *VIF* = 1.32; Distance, *VIF* = 1.31).

A likelihood ratio test of the model with the effects of boundedness and distance against a model without these effects showed a significant difference between the models (χ^2 (3) = 102.83, $p < 0.001$). Input from the best-fitting model (the one with both properties) indicates that, as in van Tiel et al. (2016), both boundedness and distance are significant predictors in deriving SIs. Specifically, the more resistant the stronger scale mate is to being perceived as an interval, the more likely it is for an SI to be triggered (logit coefficient: 0.74, *SE* = 0.12, $z = 5.9$, $p < 0.001$; Table 2). In addition, the more two items are perceived as distant on a scale (i.e., as non-interchangeable), the more likely it is for an SI to be triggered (logit coefficient: 1.41, *SE* = 0.14, $z = 9.86$, $p < 0.001$). Notably, these two factors accounted for more of the explained variance than in van Tiel et al. (49% in our study; 22% in van Tiel et al., 2016).

We performed a further analysis to test the individual contribution of each predictor, similarly to van Tiel et al., using the PartR2 package (Stoffel et al., 2020). This analysis indicated that boundedness accounted for 14.5% of the explained variance, a modest improvement from van Tiel et al., where boundedness explained 10%. Distance scores, however, accounted for 36% of the explained variance in our study, compared with 3% in van Tiel et al.'s original study. Finally, we measured the effect size of each of these predictors using the effsize package (Torchiano, 2020). There was a medium effect size for boundedness (=0.40) and a strong one for distance (=0.78).

The model also revealed a two-way interaction between our measurement of boundedness and distance ($p < 0.001$). We used the Simple Slopes analysis from the interactions package to explore it (Long, 2019). The analysis indicates that in high distance scores (1 *SD* above the mean), increasing levels of boundedness do not significantly increase the odds of SI derivation (est.: 0.14, *SE* = 0.15, $z = 0.91$, $p = 0.36$; Figure 9). However, in mean distance scores, increasing levels of boundedness significantly increase the odds of SI derivation (est.: 0.74, *SE* = 0.13, $z = 5.90$, $p < 0.001$; Figure 9), and even more so for low distance scores (1 *SD* below the mean) (est.: 1.32, *SE* = 0.19, $z = 7.10$, $p < 0.001$; Figure 9).

https://doi.org/10.1017/langcog.2024.36 Published online by Cambridge University Press

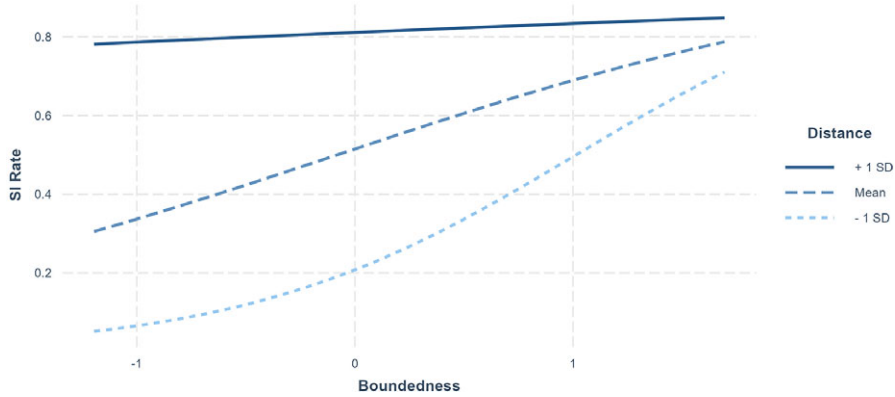

Figure 9. The influence of the interaction between boundedness and distance on SI rates. The *y*-axis represents the likelihood of deriving an SI. The *x*-axis represents boundedness scores in *SD*-units. Each line on the graph represents distance scores in *SD*-units. Thus, for example, the uppermost line stands for scales that received distance scores 1sd higher than the mean. The boundedness score of these scales did not significantly contribute to the likelihood of deriving an SI (which was already high). However, boundedness scores for scales with mean distance scores (middle line) and 1 *SD* below the mean distance scores (bottom line) were significantly affected by the boundedness score of the relevant scale.

In sum, treating boundedness as a bias and investigating distance without presupposing scalarity allowed us to corroborate and even strengthen van Tiel et al.'s original observation of an inconsistent inferential pattern for SIs that can be predicted based on the structural properties of the relevant scale. However, in our analysis, the pattern of interaction, the coefficients of both predictors, and the effect size of each predictor all show that distance is more prominent than boundedness for SI derivation.

3.2. Cluster analysis

Our second analysis was driven by suggestions in the literature that the distribution of SI rates may consist of two categories, with some distinguishing between whether the scale is adjectival or not (Beltrama & Xiang, 2013; Doran et al., 2009; Gotzner et al., 2018), and others distinguishing between properties of the scale's structure (e.g., the two categories of L-scales and M-scales as suggested in Benz et al., 2018). To explore these suggestions, we applied a cluster analysis. In cluster analyses, sets of items are grouped in such a way that items within the same cluster exhibit greater similarity to each other vis-a-vis some properties than they do to items in the other cluster(s). In the clustering of our data, we specifically followed the suggestion that the properties of the scale's structure constitute the factor underlying the derivation of SIs. Accordingly, clusters were formed based on boundedness (Experiment 2) and distance (Experiment 3), and they did not include the SI rates gathered in Experiment 1-a.

A cluster analysis involves two steps: establishing the number of clusters and partitioning the data. In the first step, we established the number of clusters in the data, the *k*-value, using both boundedness and distance scores as clustering

https://doi.org/10.1017/langcog.2024.36 Published online by Cambridge University Press

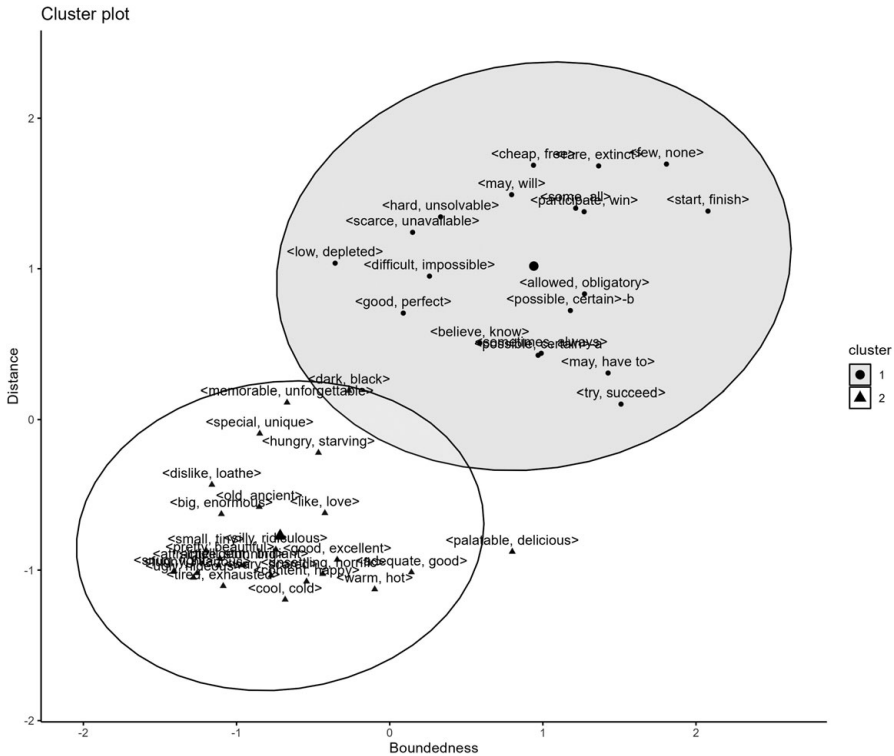

Figure 10. A *k*-means cluster analysis using boundedness and distance scores as the clustering variables. Cluster 1 is represented by the higher, darker cluster, whereas Cluster 2 is represented by the lower, lighter cluster. The axes represent the standardized boundedness scores (*x*-axis) and distance scores (*y*-axis).

variables (by data, we mean the 44 experimental scales). We used the factoextra package (Kassambara & Mundt, 2020) to derive this value by applying three different methods: Within-cluster-Sum of Squared (WSS) errors, Silhouette, and the Gap statistics (Tibshirani et al., 2001). All methods indicated that k=2 is the ideal number. It is important to note that in cluster analyses, the k-value is optimal when the objective methods converge on the same value, and this value is also theoretically motivated, as discussed above. The second step involved partitioning the n scales (=44) into our k-value (=2). To do so, we carried out a k-means cluster analysis and visualized it using the ggfortify package (Tang et al., 2016) (Figure 10 and Table 3).

Cluster 1, the dark grey cluster, consists of 19 scales, listed in Table 3. These scales received higher-than-average boundedness and distance scores. Cluster 2, the light grey cluster, consists of 25 scales, listed in Table 3. These scales received lower-than-average boundedness and distance scores. As stated, our goal was to examine whether SI patterns follow boundedness and distance patterns. Interestingly, when we examine SI rates—which were *not* included in the clustering process—we see that the average SI rate in Cluster 1 is 80%, but only 19% in Cluster 2 (i.e., the percentage of 'Yes' responses).

https://doi.org/10.1017/langcog.2024.36 Published online by Cambridge University Press

Table 3. Items categorized based on the k-means cluster analysis. Cluster 1 is referred to as Given-scales, while Cluster 2 is referred to as Volatile-scales (we discuss this terminology in the General Discussion)

Given-scales (Cluster 1) (Cluster's center: Boundedness = 1.01*SD*; Distance = 0.94*SD*)	Volatile-scales (Cluster 2) (Cluster's center: Boundedness = -0.77*SD*; Distance = -0.71*SD*)
<*some*, *all*>	<*dislike*, *loathe*>
<*few*, *none*>	<*like*, *love*>
<*sometimes*, *always*>	<*adequate*, *good*>
<*may*, *have to*>	<*attractive*, *stunning*>
<*may*, *will*>	<*big*, *enormous*>
<*believe*, *know*>	<*content*, *happy*>
<*participate*, *win*>	<*cool*, *cold*>
<*start*, *finish*>	<*dark*, *black*>
<*try*, *succeed*>	<*funny*, *hilarious*>
<*allowed*, *obligatory*>	<*good*, *excellent*>
<*cheap*, *free*>	<*hungry*, *starving*>
<*difficult*, *impossible*>	<*intelligent*, *brilliant*>
<*good*, *perfect*>	<*memorable*, *unforgettable*>
<*hard*, *unsolvable*>	<*old*, *ancient*>
<*low*, *depleted*>	<*palatable*, *delicious*>
<*possible*, *certain*>$_{-a}$	<*pretty*, *beautiful*>
<*rare*, *extinct*>	<*silly*, *ridiculous*>
<*scarce*, *unavailable*>	<*small*, *tiny*>
<*possible*, *certain*>$_{-b}$	<*snug*, *tight*>
	<*special*, *unique*>
	<*tired*, *exhausted*>
	<*ugly*, *hideous*>
	<*unsettling*, *horrific*>
	<*warm*, *hot*>
	<*wary*, *scared*>

This analysis offers a more nuanced conclusion than the regression analysis. When considering the option of two datasets within the data, we receive two different inferential patterns. In one subset, SIs are pretty robust; in fact, in 13 out of 19 scales in cluster 1, an SI was derived more than 80% of the time. In the other subset, SIs are triggered inconsistently. There, in 23 out of 25 scales in cluster 2, an SI was derived less than 50% of the time.

Interestingly, ~70% (23/33) of the adjectives were clustered together in one cluster (Cluster 2). Given this majority, one might take this finding to support the assumptions in the literature that adjectives form a distinct class of scales (Beltrama & Xiang, 2013; Doran et al., 2009; Gotzner et al., 2018). However, the sorting principle we intend to put forth is not related to the grammatical category of scales (whether the scale is adjectival or not). Instead, we suggest that the salience of the scalar construal, as determined by the scale's structural properties, is the factor that distinguishes scales in Cluster 1 from those in Cluster 2 (Section 4).

4. General discussion

We performed four experiments to examine scalar diversity and the sources of its diversity. Experiment 1-a was a slightly modified replication of van Tiel et al.'s Experiment 2 (2016), which successfully replicated the scalar diversity pattern

https://doi.org/10.1017/langcog.2024.36 Published online by Cambridge University Press

observed there. Experiment 1-b showed it to be unlikely that negative strengthening interfered with the SI rates obtained in Experiment 1-a. Experiments 2 and 3 further measured the two structural properties that van Tiel et al. found to be relevant to SI derivation: boundedness and distance. We measured boundedness as a bias using comparative constructions (Experiment 2) and distance without the presupposition of strength, and hence scalarity, using an interchangeability task (Experiment 3). Two general analyses were performed to explore the combined data: traditional regression analysis and complementary cluster analysis. We first discuss the contribution of the methodological changes (Section 4.1), and then explore the theoretical implications of this study (Section 4.2).

4.1. Methodological contribution

The current study generally followed the work of van Tiel et al. (2016) on scalar diversity. However, it offered a novel empirical investigation of boundedness and distance. By treating boundedness as a bias and by measuring distance as a function of interchangeability, we corroborated and even strengthened the results of van Tiel et al. (2016). In this section, we discuss the methodological contribution of these changes.

Overall, our empirical examination of boundedness generally supported the intuitions assumed in the literature. Scalar expressions considered bounded by van Tiel et al. and mostly perceived as such in the literature, received scores higher than the median score, and vice versa. While this pattern might, therefore, seem to support treating boundedness as an either-or property, the expressions considered strictly bounded in the literature revealed a considerable range, with values ranging from 38.1 for 'unique' to 92.97 for 'none'. These results suggest that the strict binary characterization of boundedness in the literature is, nevertheless, too rigid. According to the prevalent approach, the measurement of boundedness should have displayed values closer to the ends of our measurement scale and with a narrower range. In effect, a different pattern of results emerges from Experiment 2. Expressions traditionally classified as bounded, with boundedness scores higher than the median, display a mean further from the edge of the maximal score (and greater variability). This suggests a flexible category, which supports our initial intuition. At the same time, expressions traditionally classified as non-bounded, with boundedness scores lower than the median, display values close to the edge of the minimal score. This seems to suggest a homogenous category, which supports the traditional categorical view of boundedness.

Taking all this into account, we maintain but refine our initial position. We propose that the data point to the treatment of boundedness as a category with internal flexibility (see also Beltrama, 2018; Paradis, 2001). That is to say, within the category of bounded expressions, these expressions demonstrate varying degrees of resistance to being construed as an interval. While some bounded expressions can easily be conceptualized as denoting an interval (e.g., 'unique'), others can hardly be interpreted as anything other than an endpoint (e.g., 'extinct'). The category of non-bounded expressions, by contrast, exhibits homogeneity. Expressions denoting an interval are resistant to being conceptualized as an endpoint. In sum, while boundedness can be considered a categorical property, the categories differ from one

https://doi.org/10.1017/langcog.2024.36 Published online by Cambridge University Press

another. Bounded expressions constitute a category with internal flexibility, whereas non-bounded expressions lack this flexibility.

Concerning the measurement of distance, two outcomes distinctly stand out. First, by merely changing the task to avoid the presupposition of distance, we increased the predictive effect of distance. This increase reversed the importance of the two predictors, making distance the prominent structural property for predicting scalar diversity. Second, the outcome that a quarter of the scales received a distance score below 25 (see the Supplementary Materials) raises doubts about the canonical set of scalar expressions, or more precisely, regarding their exact nature. To reiterate, if two scalar expressions are perceived as interchangeable and, hence, as having a low/-minimal distance score, this indicates that they are perceived as denoting a similar interval (on the same scale). We discuss this further in Section 4.2.

Before turning to the theoretical implications of our findings, some task effects should be considered. First, it is possible that using the term 'interchangeable' prompted participants to perceive the two scale mates as having no distance between them. This would then undermine the conceptualization of the scale, which requires some distance between scale mates. In view of the diversity in distance scores, we can assert that this is not a core effect of the task. In fact, we would like to stress that differences between distance scores in our task and that of van Tiel et al. were limited to a subset of scalar expressions (see the next section). That is to say, if the effect were associated with the task, it would have affected all scales to the same degree. We, therefore, posit that this task effect is more reflective of the nature of these items than the nature of the task. Second, the label 'not interchangeable' in the task could potentially be selected to indicate that the relevant expressions have a maximal distance even though they are not on the same scale. Yet we used the same expressions as van Tiel et al. (2016), which are paired expressions on the same scale and for which scalarity was never questioned. Therefore, we contend that maximal distance scores for the paired expressions in this study are unlikely to represent expressions from different Horn/lexical scales. Finally, it is possible that participants understood the task as asking to what extent the two expressions can appear in the same sentential context rather than whether they have the same meaning in the same context. Since all expressions tested in this study can, in principle, appear in the same sentential context, the diversity in distance scores suggests that the task was interpreted as asking about the same meaning and not the same sentential context.

4.2. Theoretical implications

4.2.1. The relationship between boundedness and distance

Van Tiel et al. (2016) considered boundedness and distance as ways to operationalize the distinctness between scale mates. According to van Tiel et al., if distinguishing between scale mates is challenging, it is less likely that an SI will be derived. The association between distinctness and distance is apparent: The greater the distance between the lower bounds of two scale mates, the easier it is to distinguish between them. As for boundedness, van Tiel et al. suggested that '[…] scalar expressions on bounded scales are *easier* to distinguish than on non-bounded scales' (p. 163, emphasis added). This accords with other suggestions in the literature (Beltrama & Xiang, 2013; Leffel et al., 2019). Based on our findings, specifically on the pattern of interaction between boundedness and distance and the prominent role played by

https://doi.org/10.1017/langcog.2024.36 Published online by Cambridge University Press

distance in our model, we can offer an explanation as to *why* boundedness makes the distinction between scalar expressions easy.[3]

The interaction between our measurements of boundedness and distance demonstrates that an SI is derived when there is a high distance score between scalar expressions regardless of the boundedness of the stronger expression. However, when the distance score is low or medium, boundedness increases SI rates. We suggest that this is because boundedness helps to fix distance. That is to say, a scale mate conceptualized as bounded imposes an absolute distance, regardless of the size of the distance. This is because in scales conceptualized as bounded, the lower bounds of the weak and strong scale mates *necessarily* do not overlap. To illustrate, consider the scale <*rare*, *extinct*>. The stronger scale mate 'extinct' denotes an endpoint, and its value (both lower and upper bounds) corresponds to 'no existence' (=0-existence). The weaker scale mate 'rare' denotes an interval. The lower bound of 'rare' corresponds to 'at least some existence' (>0-existence; more than zero). Clearly, there is no overlap between the lower bounds of the two expressions (the points between which distance is measured), making the distance necessary and absolute. Importantly, by imposing an absolute distance, we also impose a scalar construal (see Section 1).

Moreover, we argue that in the absence of a structural property that fixes a distance, as for non-bounded scales, distance is more unstable. To support this argument, we integrate two independent observations. First, while overall, our distance scores (Experiment 3) correlated with the distance scores in van Tiel et al.'s study (see Section 3.1), a closer examination reveals a difference between bounded and non-bounded scales. The distance scores of non-bounded scales, i.e., scales that received a boundedness score below the median score, do not correlate with the distance scores in van Tiel et al. ($r = 0.27$, $p = 0.28$). However, the distance scores of bounded scales, i.e., scales that received a boundedness score above the median score, do correlate with distance scores in van Tiel et al. ($r = 0.5$, $p < 0.05$). In other words, the distance scores of the non-bounded scales fluctuated between the two experiments: many of the non-bounded scales that were given low distance scores in our study were perceived as distant in van Tiel et al.'s study. Although this may initially appear to suggest an inconsistency between the results of the two experiments—ours and that of van Tiel et al. (2016)—we propose that it actually mirrors the volatile nature of distance between expressions in non-bounded scales. Put differently, regardless of the suitability of one task over the other for measuring distance, the observed differential effect is itself informative, supporting our proposal that the distance between *some* scalar expressions, on the same scale, is more volatile.

As a second observation in support of our proposal—that in the absence of a structural property that fixes the distance between scale mates, distance is unstable—we consider the only different factor between the measurement of distance in our study and van Tiel et al.'s study: the task. The two tasks posed different contextual manipulations, or highlighted different Questions Under Discussion (QUDs) (Roberts, 1996). Based on the fact that the two tasks yielded differential results, we suggest that QUD, like boundedness, plays a role in forming distance (also see

[3]It has been suggested that alternatives are more salient with bounded expressions (e.g., Frazier et al., 2008; Gotzner et al., 2018; Van Tiel et al., 2016). However, the salience of alternatives does not appear to predict SI Diversity. For example, van Tiel et al.'s measures of 'availability', which reflect salience, did not predict SI diversity. Consequently, we followed and contributed to the idea of 'distinctness', rather than 'availability'/'salience', as the more plausible explanation as to why end-point-denoting expressions elicit more SIs.

https://doi.org/10.1017/langcog.2024.36 Published online by Cambridge University Press

Ronai & Xiang, 2022). The QUD can promote a construal of the distance as negligible (when asked about interchangeability, as in our Experiment 3), but, alternatively, it can promote a construal of a significant distance (when asked about strength, as in van Tiel et al., 2016). Notably, the QUD affected the distance scores in a subset of the scales. More specifically, while it did not (significantly) affect the distance scores between expressions in bounded scales, it did have an effect on the distance scores between expressions in non-bounded scales (as discussed above).[4] We take this task effect to reflect the nature of non-bounded expressions rather than to be inherent to the task itself. In other words, this effect indicates that the distance between adjacent scale mates in non-bounded scales is not sufficiently distinctive/stable to remain consistent across contexts because nothing in the structure of these scales forces it to do so.

In sum, our results highlight the three factors that prompt distinctness: (1) large distance (as discussed in van Tiel et al. (2016) and demonstrated here as well); (2) a conceptually fixed/absolute distance imposed by a structural property (as developed here with respect to boundedness); and (3) contextual manipulations (see, e.g., Doran et al., 2012 and Ronai & Xiang, 2022 for the role of the QUD in triggering SIs, or Breheny et al., 2006 for the effect of different contexts in triggering SIs).[5]

4.2.2. *The relationship between scales and inferences*

Our cluster analysis suggested that the data consists of two clusters, based on boundedness and distance scores. To explain the sorting principle that distinguishes items in Cluster 1 from items in Cluster 2, we here develop the suggestion proposed above: boundedness imposes an absolute distance.

Cluster 1 comprises scales that receive higher-than-average distance and boundedness scores, in other words, scales in which distance is large or fixed, hence the scalar construal of these scale mates is entrenched. To emphasize, a large distance between two scale mates results in an entrenched scalar construal (regardless of boundedness). Moreover, even when the distance between two scale mates is low or medium, an entrenched scalar construal occurs when the scale is bounded (see Section 4.2.1). Scales with an entrenched scalar construal can be conceived of as Given-scales (in the sense suggested in Gazdar, 1979, p. 58). Cluster 2 primarily comprises scales that received lower-than-average boundedness and distance scores. These properties, and possibly others not addressed in this study, contribute to the fact that their scalar construal fluctuates. We refer to these scales as Volatile-scales.

[4]This idea aligns nicely with the literature on adjectives, specifically with the idea that the standard of comparison is fixed for absolute gradable adjectives (what we broadly refer to here as bounded expressions, e.g., 'extinct'), but context-dependent for relative gradable adjectives (what we refer to here as non-bounded expressions, e.g., 'stunning'; see Kennedy & McNally, 2005; Kennedy, 2007; for empirical findings, see Frazier et al., 2008; Gotzner et al., 2018). It is possible that the QUD effect on the perceived distance between two expressions is limited when the interpretation of the relevant expressions is (relatively) resistant to context variations.

[5]We note that this is not meant to serve as an exhaustive list of factors affecting scalar diversity or distinctness. As an example of other factors not tested here, let us consider Gotzner et al. (2018), who observed that negative adjectives denoting the complete absence of some quantity (e.g., 'extinct', 'free') yielded the highest SI rates (Gotzner et al., 2018, p. 11). Applying the same reasoning to these adjectives as for boundedness, we propose that the absence of the underlying property can also impose an absolute distance and prompt the derivation of SIs.

https://doi.org/10.1017/langcog.2024.36 Published online by Cambridge University Press

Thus, while previous suggestions focused on grammatical categories as a sorting principle, separating adjectival scales from the other scales (Beltrama & Xiang, 2013; Doran et al., 2009; Gotzner et al., 2018), and although Cluster 2 is primarily comprised of adjectival scales, we propose that what essentially sets the items in Cluster 2 apart from those in Cluster 1 is a structural property, namely, distance. The distance in Given-scales (Cluster 1) is conceptualized as entrenched, while in Volatile-scales (Cluster 2) it is fluctuant.

Our perspective here diverges from the standard view, thus necessitating clarification. We distinguish between the scalar construal of Given- and Volatile-scales. Given-scales have a rigid scalar construal, due to the large or even fixed distance between the relevant paired expressions. However, we propose that paired expressions on Volatile-scales can be conceptualized either as scalars or possibly even as near-synonyms, in light of the instability of distance between them. In the latter case, the construal one adopts depends on various factors, such as the QUD, individual differences, language-specific considerations, etc. In other words, it is context-dependent. Importantly, the boundary between Given- and Volatile-scales is not rigid. It is possible that some items are harder to classify into one category or another. This may relate to how different individuals perceive distance. In other words, it is important to avoid fixating on assigning specific items to particular clusters while acknowledging the existence of two scale types.

When considering the inferential patterns associated with each of these types of scales, it is not surprising that Given-scales, with a fixed distance, and hence an entrenched scalar construal, trigger SIs robustly (~80% on average), whereas Volatile-scales, with a fluctuant distance, and hence an unstable scalar construal, trigger SIs inconsistently (only ~20% on average). If the two scale mates are not perceived as sufficiently distinct, there is no reason to reject the 'stronger' scale mate. For instance, if the participant perceived 'snug' and 'tight' as close in meaning, there is no reason to reject 'tight' when 'snug' is expressed. Notably, it follows that the fundamental trigger for scalar inferences is the ability to conceptualize a scale, that is, the ability to view the two scale mates as distinct. After all, you cannot have a scalar inference if you do not *first* have a scale.

4.2.3. The coexistence of scalar diversity and SI uniformity

Previous research on scalar diversity extended the set of scales to achieve a better understanding of the mechanism underlying SIs. Consequently, when faced with SI diversity, this diversity was attributed to the inferential mechanism, which could no longer be regarded as consistent or uniform (but see Benz et al., 2018; Gotzner et al., 2018). While the findings in this study affirm the diversity of SIs (Figure 3), we suggest an alternative explanation to account for it. Specifically, we suggest that by acknowledging that some scalar construals are entrenched (those we refer to as Given-scales), while others are fluctuant (those we refer to as Volatile-scales), a different, more plausible explanation for SI diversity arises.

Considering that scalar inferences involve the rejection of a stronger alternative, it is clear why statements including 'Volatile' scalar expressions do not necessarily trigger an SI. When the alternative scalar expression is perceived as equivalently strong, that is, as denoting a close meaning, there is no reason to reject it. This is unlikely to happen in statements that include 'Given' scalar expressions, where the construal forces a difference in strength/distance. Accordingly, a fundamental trigger

https://doi.org/10.1017/langcog.2024.36 Published online by Cambridge University Press

for scalar inferences is the ability to first conceptualize a scale, that is, a difference in distance between the (lower bounds of the) two scalar expressions. Once a scalar construal has been conceptualized, even for Volatile scales, one can assume that an SI will be derived, that is, the stronger alternative will be rejected.

If this is so, we propose that SI diversity arises from the fluctuant scalar construal of certain scales. When two scale mates are not easily distinguishable, an SI will be triggered only when a scale is construed. However, once there is a scale, either intrinsically through structure (as in Given-scales) or extrinsically through context (as might happen in Volatile-scales), it is likely that an SI will be derived using the same processing procedure, of stronger alternative rejection. In other words, diversity arises from the process of conceptualizing the scale and not from the derivation of the inference. Currently, the existing literature on SI diversity sets out from the assumption that two scalar expressions, at least the items in this study, invariably form a scale. We argue that some of these 'scale mates' are not always construed as forming a scale. We therefore impose a preliminary step for SI derivation: namely, a scale must first be construed.

The idea that constructing a scale is essential for deriving an SI has been raised in the literature concerning the development of this linguistic ability. Several studies using various tasks have shown that children succeed in deriving SIs using numbers (i.e., arriving at the exact meaning) but not quantifiers (Huang & Snedeker, 2009; Hurewitz et al., 2006; Noveck, 2001; Papafragou & Musolino, 2003). One explanation for this dissociation is attributed to differences in the ease of scale conceptualization (Barner et al., 2011). For numerals, children find it easy to conceptualize a scale because they are well-familiar with the number line (at least in Western societies). However, when it comes to quantifiers, children struggle to establish the connection between 'some' and 'all' and, hence, have difficulties forming the scale, hindering their access to the strong alternative. This is compatible with our proposal that the derivation of SIs necessarily builds on the preliminary step of construing the scale.

For the current discussion, we wish to stress that by acknowledging the preliminary step of scale construal, we can, in fact, account for the *diversity across scales* based on whether or not a scalar construal was conceptualized, while maintaining the *uniformity* of the inferential mechanism *for all scales* once this scalar construal was conceptualized.

5. Conclusion

This study follows in the footsteps of previous work on scalar diversity but goes beyond them, proposing new methodological and theoretical implications. Methodologically, we applied different measures to investigate two of the structural properties that predict the likelihood of deriving SIs, namely boundedness and distance. In so doing, we reinforced the previous results attained by van Tiel et al. (2016) and offered new insights into these predictors: We demonstrated why treating boundedness as an absolute property is too strict, suggesting that it should instead be treated as a category with internal flexibility (following Paradis, 2001). Furthermore, we demonstrated the problem of presupposing distance when examining (canonical) scalar expressions. In so doing, we were able to highlight the crucial role of distance in SI derivation. We examined the evidence using

https://doi.org/10.1017/langcog.2024.36 Published online by Cambridge University Press

traditional regression analysis as well as complementary cluster analysis. We proposed why and how boundedness can fix distance, making it absolute. Moreover, we proposed that fixing distance is one of at least three possible ways to achieve distinctness between scale mates on some scale: large distance, fixed/absolute distance, and contextual manipulation. We referred to structurally-distinct scalar expressions as Given-scales and to those characterized by a fluctuant structure as Volatile-scales. We emphasized that these structural differences are indicative of two distinct scale types, associated with two distinct inferential patterns: Given-scales with an entrenched scalar construal trigger SIs robustly, whereas Volatile-scales with a fluctuant scalar construal trigger SIs inconsistently. Based on these patterns, we suggested that SI diversity arises from difficulties in conceptualizing a scale: a precondition for the derivation of SIs. We consequently argued that by acknowledging this precondition, we can account for SI diversity, which depends on whether a scale is construed, while maintaining the assumption concerning the uniformity of the derivational process once the scale has been construed.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/langcog.2024.36.

Acknowledgments. Our sincere thanks to all members of the Language in Social Context laboratory for their helpful comments and discussions, as well as special thanks to Nitzan Trainin for his invaluable support. We also wish to express our appreciation to our two anonymous reviewers for their insightful and constructive feedback.

Funding. This study was supported by the Israel Science Foundation (ISF) (grant no. 1398/20), awarded to Mira Ariel, and the Israel Science Foundation (ISF) (grant no. 811/23), awarded to Einat Shetreet.

Competing interest. The authors declare none.

References

Ariel, M. (2004). Most. *Language*, 80, 658–706.

Ariel, M. (2015). Doubling up: Two upper bounds for scalars. *Linguistics*, 53(3), 561–610. https://doi.org/10.1515/ling-2015-0013

Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition*, 118(1), 84–93. https://doi.org/10.1016/j.cognition.2010.10.010

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beltrama, A. (2018). Totally between subjectivity and discourse. Exploring the pragmatic side of intensification. *Journal of Semantics*, 35(2), 219–261. https://doi.org/10.1093/semant/ffx021

Beltrama, A., & Xiang, M. (2013). Is excellent better than good? Adjective scales and scalar implicatures. *Sinn Und Bedeutung*, 17, 81–98.

Benz, A., Bombi, C., & Gotzner, N. (2018). Scalar diversity and negative strengthening. *ZAS Papers in Linguistics*, 60, 191–203. https://doi.org/10.21248/zaspil.60.2018.462

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

Devlesschouwer, L. (2019). Upper-bounded scalars and argumentation-in-language theory. *Anglophonia*, 28. https://doi.org/10.4000/anglophonia.2580

Dillon, B., & Wagers, M. W. (2021). Approaching gradience in acceptability with the tools of signal detection theory. In G. Goodall (Ed.), *The Cambridge handbook of experimental syntax* (1st ed., pp. 62–96). Cambridge University Press. https://doi.org/10.1017/9781108569620.004

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(2), 211–248. https://doi.org/10.1163/187730909X12538045489854

Doran, R., Ward, G., Larson, M., McNabb, Y., & Baker, R. E. (2012). A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language*, 88(1), 124–154. https://doi.org/10.1353/lan.2012.0008

Fox, J., & Weisberg, S. (2019). Companion to applied regression (Version 3) [Computer software]. Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Frazier, L., Clifton, C., & Stolterfoht, B. (2008). Scale structure: Processing minimum standard and maximum standard scalar adjectives. *Cognition*, 106(1), 299–324. https://doi.org/10.1016/j.cognition.2007.02.004

Gazdar, G. (1979). *Pragmatics: Implicature, presupposition, and logical form*. Academic press.

Geurts, B. (2010). *Quantity implicatures* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511975158

Gotzner, N., Solt, S., & Benz, A. (2018). Scalar diversity, negative strengthening, and adjectival semantics. *Frontiers in Psychology*, 9, 1659. https://doi.org/10.3389/fpsyg.2018.01659

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics* (Vol. 3, pp. 41–58). Academic Press.

Grice, H. P. (1989). *Studies in the way of words*. Harvard University Press.

Hirschberg, J. (1985). *A theory of scalar implicature. University of Pennsylvania.*

Horn, L. R. (1972). *On the semantic properties of logical operators in English*. University of California.

Horn, L. R. (1989). *A natural history of negation*. The University of Chicago Press.

Huang, Y. T., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive Psychology*, 58(3), 376–415. https://doi.org/10.1016/j.cogpsych.2008.09.001

Hurewitz, F., Papafragou, A., Gleitman, L., & Gelman, R. (2006). Asymmetries in the acquisition of numbers and quantifiers. *Language Learning and Development*, 2(2), 77–96. https://doi.org/10.1207/s15473341lld0202_1

Kanai, R., Walsh, V., & Tseng, C. (2010). Subjective discriminability of invisibility: A framework for distinguishing perceptual and attentional failures of awareness. *Consciousness and Cognition*, 19(4), 1045–1057. https://doi.org/10.1016/j.concog.2010.06.003

Kassambara, A., & Mundt, F. (2020). Factoextra: Extract and visualize the results of multivariate data analyses (R Package Version 1.0.7.) [Computer software]. https://CRAN.R-project.org/package=factoextra.

Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, 30(1), 1–45. https://doi.org/10.1007/s10988-006-9008-0

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

Koenig, J.-P. (1991). Scalar predicates and negation: Punctual semantics and interval interpretations. In L. Dobrin, L. Nichols, & R. Rodriguez (Eds.), *Proceedings of CLS 27* (pp. 140–155). Chicago Linguistic Society.

Leffel, T., Cremers, A., Gotzner, N., & Romoli, J. (2019). Vagueness in implicature: The case of modified adjectives. *Journal of Semantics*, 36(2), 317–348. https://doi.org/10.1093/jos/ffy020

Long, J. (2019). Interactions: Comprehensive, user-friendly toolkit for probing interactions [Computer software]. https://cran.r-project.org/package=interactions.

Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, 78(2), 165–188. https://doi.org/10.1016/S0010-0277(00)00114-1

Orr, S., Ariel, M., & Peleg, O. (2017). The case of literally true propositions with false implicatures. In I. Chiluwa (Ed.), *Deception and deceptive communication: Motivations, recognition techniques and behavioral control* (pp. 67–107). Nova Science Publishers, Inc.

Orr, S., Ariel, M., & Shetreet, E. (2023). Scalar structures cannot impose semantic meaning. In *The 10th biennial meeting of Experimental Pragmatics, Paris*.

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Pankratz, E., & Van Tiel, B. (2021). The role of relevance for scalar diversity: A usage-based approach. *Language and Cognition*, 13(4), 562–594. https://doi.org/10.1017/langcog.2021.13

Papafragou, A., & Musolino, J. (2003). Scalar implicatures: Experiments at the semantics–pragmatics interface. *Cognition*, 86(3), 253–282. https://doi.org/10.1016/S0010-0277(02)00179-8

Paradis, C. (2001). Adjectives and boundedness. *Cognitive Linguistics*, 12(1), 47–64. https://doi.org/10.1515/cogl.12.1.47

R Core Team. (2021). R: A language and environment for statistical computing [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/.

Rizopoulos, D. (2006). ltm: An *R* Package for Latent Variable Modeling and Item Response Theory Analyses. *Journal of Statistical Software*, 17(5). https://doi.org/10.18637/jss.v017.i05

Roberts, C. (1996). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5.

Ronai, E., & Xiang, M. (2022). Quantifying semantic and pragmatic effects on scalar diversity. *Proceedings of the Linguistic Society of America*, 7(1), 5216. https://doi.org/10.3765/plsa.v7i1.5216

RStudio Team. (2020). RStudio: Integrated Development for R. RStudio [Computer software]. RStudio, PBC. http://www.rstudio.com/.

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

Soames, S. (1982). How presuppositions are inherited: A solution to the projection problem. *Linguistic Inquiry*, 13, 483–545.

Sternau, M., Ariel, M., Giora, R., & Fein, O. (2015). Levels of interpretation: New tools for characterizing intended meanings. *Journal of Pragmatics*, 84, 86–101. https://doi.org/10.1016/j.pragma.2015.05.002

Stoffel, M. A., Nakagawa, S., & Schielzeth, H. (2020). *partR2: Partitioning R 2 in generalized linear mixed models [Preprint]. Bioinformatics.* https://doi.org/10.1101/2020.07.26.221168

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

Tang, Y., Horikoshi, M., & Li, W. (2016). ggfortify: Unified interface to visualize statistical results of popular R packages. *The R Journal*, 8(2), 474. https://doi.org/10.32614/RJ-2016-060

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2), 411–423. https://doi.org/10.1111/1467-9868.00293

Torchiano, M. (2020). effsize: Efficient effect size computation [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.1480624

Van Rooij, R., & Schulz, K. (2004). Exhaustive interpretation of complex sentences. *Journal of Logic, Language and Information*, 13(4), 491–519. https://doi.org/10.1007/s10849-004-2118-6

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

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), 1686. https://doi.org/10.21105/joss.01686

Cite this article: Orr, S., Ariel, M., & Shetreet, E. (2024). Scales and inferences, *Language and Cognition*, 1–26. https://doi.org/10.1017/langcog.2024.36