On the Non-Unified Nature of Scalar Implicature: An Empirical Investigation

Ryan Doran  
Northwestern University, USA  
r-doran@northwestern.edu

Rachel E. Baker  
Northwestern University, USA  
relibaker@yahoo.com

Yaron McNabb  
University of Chicago, USA  
ymcnabb@uchicago.edu

Meredith Larson  
Northwestern University, USA  
meredithjlarson@gmail.com

Gregory Ward  
Northwestern University, USA  
gw@northwestern.edu

Abstract
Scalar implicature is often offered as the exemplar of generalized conversational implicature. However, despite the wealth of literature devoted to both the phenomenon in general and to specific examples, little attention has been paid to the various factors that may influence the generation and interpretation of scalar implicatures. This study employs the “Literal Lucy” methodology developed in Larson et al. (in press) to further investigate these factors in a controlled experimental setting. The results of our empirical investigation suggest that the type of scale employed affects whether or not speakers judge a particular scalar implicature to be part of the truth-conditional meaning of an utterance. Moreover, we found that features of the conversational context in which the implicature is situated also play an important role. Specifically, we have found that the number of scalar values evoked in the discourse context plays a significant role in the interpretation of scalar implicatures generated from gradable adjective scales but not other scale types. With respect to the effects of scale type, we have found that gradable adjectives were less frequently incorporated into truth-conditional meaning than cardinals, quantificational items, and ranked orderings. Additionally, ranked orderings were incorporated less than cardinals. Thus, the results from the current study show that the interpretation of scalar implicature is sensitive to both the associated scale type and discourse context.
Introduction

Scalar implicature is the most frequently discussed type of conversational implicature, owing in large part to the regularities that hold across all examples of the phenomenon. However, despite the wealth of literature devoted to this particular type of implicature, little attention has been paid to the various factors that may influence the generation and interpretation of scalar implicatures. This study employs the methodology developed in Larson et al. (in press) to investigate the differences they found across various types of scalar implicatures. Larson et al. concluded that speakers do not treat all scalar implicatures alike with respect to their inclusion in truth-conditional meaning. For example, implicatures associated with gradable adjectives were less frequently judged to affect truth-conditional meaning than those that enrich cardinal numbers from an “at least” to an “exactly” reading. The current study presents empirical evidence that further investigates and extends these findings, showing that not only does the type of scale affect the interpretation of scalar implicatures but also that certain contextual features affect whether or not speakers judge a particular scalar implicature to be part of the truth-conditional meaning of an utterance.

1. Generalized Conversational Implicatures

Within Grice’s framework (Grice, 1967), a speaker’s communicated meaning consists of what is said and what is implicated. What is said corresponds to the truth-evaluable propositional content of an utterance, while what is implicated corresponds to additional aspects of speaker-intended meaning beyond what is said. Conversational implicatures arise from the assumption that speakers are behaving cooperatively by attempting to further the ends of the discourse through their co-participation. Speakers succeed in communicating more than what they literally say through the observation and violation of conversational maxims, i.e. ways of adhering to Grice’s Cooperative Principle. Generalized conversational implicatures (GCI) are calculable according to these conversational maxims and hence are not purely conventional but nonetheless exhibit a high degree of regularity. GCI are general in the sense that they arise under normal circumstances unless something in the context prevents or “blocks” their generation.
As additional, pragmatically-conveyed meaning, conversational implicatures for Grice may be cancelled. That is, since conversational implicatures are by definition not part of truth-conditional meaning, they may be denied by the speaker in a subsequent remark without contradiction (e.g. “but I don’t mean to suggest that…”). The cancellability feature that Grice attributed to all conversational implicature has been challenged by a number of authors. These “Post-Gricean” authors (e.g. Sperber & Wilson, 1986; Recanati, 1993; Carston, 2002) have argued that many of the examples classified as GCIs by Grice and “Neo-Griceans” (e.g. Horn, 1984; Levinson, 2000) are in fact not cancellable and, hence, are part of truth-conditional meaning. While they maintain a distinction between saying and implicating, Post-Griceans reject the Gricean claim that what is said is determined by pre-pragmatic processes. For them, there are pragmatically-determined aspects of truth-conditional meaning extending well beyond the few examples recognized by Grice, such as indexical and pronoun resolution. The issue here concerns where to draw the line between saying and implicating. For Grice, what is said is determined by the conventional meaning of the words a speaker utters, while what is implicated is determined by additional pragmatic factors. The Post-Griceans, on the other hand, deny that there is a level of meaning corresponding to what is said that is pre-pragmatic. Instead, under this view, pragmatics “intrudes” upon truth-conditional meaning, rendering some of Grice’s GCIs truth-conditional and not in fact genuine implicatures.¹

In a Gricean framework, scalar implicatures are all derived from Grice’s first Maxim of Quantity: “Make your contribution as informative as required” (Grice, 1967: 26). Through observing this maxim, a speaker may implicate that additional information is false (or unknown) because they have not provided a more informative response.² The notion of information employed by this maxim can be understood as an ordering of values on a scale, which ranks the informativeness of values in a descending order. An utterance employing a stronger value on such a scale provides more information than a similar utterance using a weaker value. Scalar implicatures are generated when a speaker selects a weaker value on the scale, thus implicating that stronger values on the scale are false. In examples (1)-(3) below, a speaker’s use of the (a) sentence will

---

¹ Our use of the term “implicature” in what follows is intended to cover all examples classified as implicatures in the literature, regardless of their ultimate theoretical status.

² Throughout this paper we will leave aside epistemic issues concerning implicature. Some authors treat the implicature that a higher scalar value is unknown as an example of scalar implicature, while others see it as a distinct type of Quality implicature or a constraint upon the generation of scalar implicatures. In the experiment presented below, these epistemic concerns are not relevant.
in normal circumstances implicate that, according to the scale in (d), the corresponding (b) sentence is false. Thus, the scalar value in each of the (a) sentences licenses the inference to the corresponding (c) sentence.

(1)  
a. Gus ate some of the cake.  
b. Gus ate all of the cake.  
c. Gus did not eat all of the cake.  
d. \( <\text{all, some}> \)

(2)  
a. She invited many of her classmates.  
b. She invited all of her classmates.  
c. She didn’t invite all of her classmates.  
d. \( <\text{all, many, a few}> \)

(3)  
a. John bought two books.  
b. John bought three books.  
c. John didn’t buy three books.  
d. \( <\ldots3, 2, 1> \)

As discussed in Horn (1972, 1984) and Gazdar (1979), scalar implicature involves the ordering of scalar values according to an entailment relation. If a sentence containing a stronger value on the scale is true, this will entail the truth of a corresponding sentence with a weaker scalar value substituted. In example (3) above, the informationally stronger claim that John bought three books entails the truth of the weaker claim that he bought two. This entailment relation orders the scale and specifies the relation between the values on the scale. Implicatures arising from modals \( <\text{must, can}> \), definites/indefinites \( <\text{the, a}> \), logical connectives \( <\text{and, or}> \), and others can be treated likewise within this framework.

Hirschberg (1991) provides an account of scalar implicatures that extends to cases in which an entailment relation doesn’t hold among scalar values. It had been previously noted that conventional rankings, such as legal classifications (tort, misdemeanor, felony, capital crime), which are not ordered on the basis of entailment relations, can also generate scalar phenomena. For example, if a company policy states that people convicted of a misdemeanor are not eligible for hire, then one is licensed to infer that people convicted of a felony are not eligible as well. Hirschberg develops the notion of a partially ordered set, or \textbf{poset}, in order to account for scalar phenomena more generally, of which entailment-based scales are but one instance. If a set of linguistic expressions can be ranked according to some metric which orders alternate values as higher or lower, a \textbf{poset} can support scalar implicatures, provided that the ranking metric is salient to both speaker and hearer. This broader notion of a partial ordering allows for not only orderings that do not support
entailments but also for orderings that are much more context- or domain-sensitive, as in (4):

(4) A: Did you get Brad Pitt’s autograph?
   B: I got Angelina Jolie’s.

   Implicature: B did not get Brad Pitt’s autograph.

Here, what makes the exchange coherent is the world knowledge that Angelina Jolie and Brad Pitt are partners and thus form a poset of individuals. Whether scales are defined by posets or by an entailment relation, however, they all share the property of having their values ordered with respect to some salient metric.

1.1. Experimental Approaches to GCIs

Previous experimental work has been largely concerned with cognitive and developmental questions about the processing of implicatures. While important for furthering our understanding of implicature phenomena, such experiments have shed little light on where to draw the line between saying and implicating. The few studies that have attempted to address whether speakers incorporate implicatures into truth-conditional meaning (Gibbs & Moise, 1997; Nicolle & Clark, 1999) have suggested that speakers often treat implicatures as affecting truth-conditional meaning. These studies, however, suffered from various methodological problems, such as instructions that assumed familiarity with technical terminology. To address these issues, Larson et al. (in press) identified an experimental paradigm in which participants were trained to interpret target items from the perspective of a literal-minded person, named Literal Lucy. This training was designed to give participants consistent criteria for the task, focussing on truth-conditional meaning rather than on what the target sentence could communicate more generally. They found that participants were less likely to judge implicatures to be part of truth-conditional meaning when using Literal Lucy’s perspective to guide their judgements.

In addition to developing a new paradigm to address where the line between saying and implicating should be drawn, Larson et al. also investigated a wider variety of GCI types than had been previously explored. Rather than focussing on a small number of standard examples, they employed stimuli that reflected a much wider range of different GCI types. On the whole, participants in these studies treated GCIs as distinct from semantic entailments as well as from other pragmatically determined elements of truth-conditional meaning, such as deixis and ellipsis. However, within the category of GCIs, there was considerable
variance in terms of the frequency with which the implicatures were incorporated into truth-conditional meaning. This effect was the most pronounced for the four types of scalar implicatures stimuli, exemplified in (5)-(8).

(5) **Cardinals**
   Irene: How many children does Lisa have?
   Sam: Lisa has three children.
   FACT: Lisa has quadruplets.

(6) **Ranked Orderings**
   Irene: Who can register for the advanced seminar?
   Sam: Juniors can register.
   FACT: Anyone who has completed his or her first year of study can register.\(^3\)

(7) **Quantifiers and Modals**
   Irene: How much cake did Gus eat at his sister’s birthday party?
   Sam: He ate most of the cake.
   FACT: By himself, Gus ate his sister’s entire birthday cake.

(8) **Gradable Adjectives**
   Irene: How attractive is Kate?
   Sam: She’s pretty.
   FACT: Kate was voted “World’s Most Beautiful Woman” this year.

Of these four types, cardinals were most frequently judged to affect truth-conditional meaning (75%), followed by ranked orderings (49%), quantifiers and modals (33%), and gradable adjectives (12%). The large degree of variance among these types of scalar implicatures was unexpected and not predicted by any current theory of scalar implicature. To draw any clear conclusions, however, a greater number of stimuli would be necessary for statistically valid comparisons across scale types.

The goal of the current study is to explore which factors might explain the observed variance in the frequency with which scalar implicatures are incorporated into truth-conditional meaning. A more frequent incorporation into truth-conditional meaning could be a result of features associated with the scale itself. As Larson et al. themselves observed, different scale types are associated with different mean rates of incorporation. For example, they found that participants were more likely to exclude stronger, yet logically consistent, scalar values for cardinals than, say, for gradable adjectives.

\(^3\) In the U.S. university system, there are typically four categories of students: freshmen (first year), sophomores (second year), juniors (third year), and seniors (fourth year).
An additional possible factor affecting the incorporation of scalar implicatures into truth-conditional meaning is the relative salience of alternate scalar values. While it has been recognized that there is a role for salience in the interpretation of scalar implicature (Hirschberg, 1991; inter alia), the factors that contribute to the salience of scalar values have not been adequately investigated. The more salient the alternate values on a scale are, the more likely a speaker may be to draw the inference that stronger values on that scale are false. A likely factor affecting the salience of scalar values is whether alternate scalar values have been explicitly evoked in context. Consider (10), in which the discourse context explicitly introduces alternate scalar values not evoked in (9).

(9) Irene: How many children does Lisa have?
   Sam: Lisa has three children.
(10) Irene: Does Lisa have three, four or five children?
   Sam: Lisa has three children.

In (10), Sam has chosen not to respond with one of the stronger scalar values evoked by Irene's question. Given that alternate values were explicitly evoked in this context, and are therefore salient, this increased salience may result in a greater likelihood that the corresponding implicature would be incorporated into truth-conditional meaning. In what follows, we describe the experiment that we designed to investigate these factors.

2. Methods

2.1. Materials

The materials used in this experiment consisted of short conversations between two characters, Irene and Sam. In these written conversations (presented on a computer screen), Irene asks Sam a question and Sam answers it. After his answer, information that is relevant to the conversation, and which the participants are told is true, appears (labelled “FACT”). In each experimental item, Sam’s response licenses a scalar implicature, and the FACT contradicts this implicature but not what is said. An example of an experimental item is provided in (11).

(11) Irene: Does Lisa have two, three, four, or more children?
   Sam: Lisa has three children.
   FACT: Lisa has quadruplets.
The experimental materials varied along two dimensions: type of scale and discourse context. Scale type was manipulated by using scales drawn from one of four semantically motivated categories (described below). Discourse context was manipulated by varying the number of scalar values that were explicitly evoked at the beginning of each experimental conversation, as part of Irene’s question.

Our four scale types included: cardinals, gradable adjectives, ranked orderings, and quantificational items. Cardinal scales consist of positive cardinal numbers (e.g. <…3, 2, 1>). Gradable adjective scales consist of adjectives that form a continuous scale of values (e.g. <…furious, upset, annoyed…>, <…sweltering, hot, warm…>). Ranked orderings consist of conventional orderings (e.g. <king, queen, full, twin>, <doctoral, master’s, bachelor’s>). Quantificational items consist of quantifiers, quantificational adverbs, and adverbial phrases (e.g. <all, most, some>, <always, frequently, sometimes>).

To investigate the effects of discourse context on participants’ judgements, we created three versions of each conversation, each of which contained a different number of values on a given scale: no values evoked (N), one value evoked (O), and multiple values evoked (M). For example, in the (N) version of an experimental conversation, Irene would evoke the relevant scale but none of the scale’s members, as in (12). In the (O) version, Irene would mention one member of the scale, as in (13), while in the (M) version, Irene would mention several members of the scale, as in (14).

(12) Irene: How many children does Lisa have? No values evoked (N)
(13) Irene: Does Lisa have four children? One value evoked (O)
(14) Irene: Does Lisa have two, three, four, or more children? Multiple values evoked (M)

In the (O) condition, Irene’s question always included a member of the scale that was stronger than that used in Sam’s response. This is illustrated in (15), where Irene uses the value four, while Sam uses the weaker value three.

(15) Irene: Does Lisa have four children?
    Sam: Lisa has three children.

We held several features of the FACT constant. First, the same FACT was used for all three versions of a given conversation. Second, all experimental FACTs

---

4 A list of all experimental items can be found in Appendix A.
5 For scales with fixed maximum and minimum values, Irene’s question in the (M) condition would use the minimum, maximum and additional intermediate values. For scales formed from infinite sets, such as cardinal numbers, Irene’s question would include three consecutive values in the appropriate range as well as an additional clause that evokes higher scalar values (as in (14)).
shared the same level of referential specificity as Irene’s question. For example, if Irene mentioned a place or person by name, this name was repeated in the FACT. However, in Sam’s response, pronouns could be used in place of full NPs. In addition, none of the FACTs contained any scalar values; instead, all of the FACTs licensed inferences to the intended value on the relevant scale. This can be seen in (16), in which the word quadruplets licenses the inference that Lisa has four children.

(16)   Irene: Does Lisa have two, three, four, or more children?
   Sam: Lisa has three children.
   FACT: Lisa has quadruplets.

The experiment had 28 control items: 14 entailments (in which the FACT entailed Sam’s statement), 14 contradictions (in which Sam’s statement contradicted the FACT), and 20 additional contextually-sensitive filler items. Examples of each of these types can be found in Appendix B.

2.2. Design

The experiment included 48 scales, each of which was used in a different conversation. For each conversation, there were three versions, one for each of the discourse context conditions ((N), (O), and (M)) described above, resulting in a total of 144 experimental items. Each participant read only one of these versions for each conversation, so each participant saw 48 experimental items in all. For example, a participant might have read the (N) version of the conversation for the <3, 2, 1> scale, the (O) version of the conversation for the <6, 5, 4> scale, the (M) version of the conversation for the <all, most, some> scale, etc. Experimental, control, and filler items were grouped using a Latin Square design, resulting in 12 experimental scripts. Each script contained 12 blocks of four experimental items and four control/filler items. The order of items within the blocks and the order of the blocks themselves were randomised, so no two participants read the same combination of conversations in the same order. In this paper, we analyse a subset of the scales included in the experiment: ten cardinals, eleven gradable adjectives, five ranked orderings, and six quantificational items, resulting in 32 scales in total.6

6 Of these, 10 were non-scalar Generalized Conversational Implicatures (GCIs) and 10 were context-sensitive elements whose interpretation is required to arrive at a truth-evaluable proposition.

7 In our original experimental design, the experimental items were divided up into slightly different categories than those analysed in this paper. Therefore, we did not include some experimental items in this analysis.
2.3. Participants

Forty-eight native speakers of North American English drawn from the Northwestern University community participated in this experiment. Thirty participants were recruited through flyers posted on campus and were paid for their participation, and 18 participants were recruited from introductory Linguistics classes, receiving course credit for their participation. No participants reported any language impairments other than one minor speech impediment, which was corrected through therapy. The mean age of our participants was 20; 14 were males, 34 were females.

2.4. Procedure

Participants were given instructions introducing them to Literal Lucy, a literal-minded character who interprets everything literally and, thus, misinterprets instances of non-literal language, such as figurative language and indirect speech acts. We used the Literal Lucy character to evoke participants’ folk notion of “interpreting literally” while avoiding relying on technical terminology. Our use of the Literal Lucy character also induced a shift in perspective that focussed participants on a sentence’s truth-conditional meaning, rather than having them rely on their own interpretation of what the speaker may have been trying to communicate.

Participants were introduced to Literal Lucy’s particular way of interpreting language through examples like (17), in which Literal Lucy demonstrates her tendency to attend only to the conventional meaning of the words uttered.

(17) Example of Literal Lucy’s interpretation of figurative language

    Frank: Brian just had a birthday, and I didn’t realize how old he was.
    Lucy: Really? How old is he?
    Frank: He just turned 40, so now he’s over the hill.
    Lucy: Hill? Which hill? And when did he go over it?

In (17), instead of interpreting the idiom over the hill idiomatically (i.e. “past one’s prime”), Literal Lucy interprets it literally (i.e. “over some particular hill”). After demonstrating how Literal Lucy interprets utterances, we instructed participants to adopt her way of interpreting language when evaluating the truth of statements in the experimental task. Each conversation was followed with “Given this FACT, Literal Lucy would say that the underlined sentence is: T or F.”, which served to remind participants to respond as Literal Lucy would. This question was followed by a four-point Likert Scale, on which participants rated their confidence in their answers. Participants performed
the tasks on a computer in a sound-attenuated booth using a self-paced reading methodology.

3. Results

3.1. Analysis

Statistical analyses of the participants’ true and false responses were carried out using linear mixed-effects logistic regression models with participants and items as random factors. This type of analysis avoids spurious effects that can arise when proportion data are analysed using traditional ANOVAs (Jaeger, 2008). In our linear regressions, when a factor (i.e. an independent variable) has multiple levels, one is selected as the reference level, and all other levels are compared to it within a model. Analyses were carried out in R, a programming language and environment for statistical computing (R Development Core Team, 2008).

3.2. Comparison of Scalar GCIs to Contradictions and Entailments

We first built two models to determine whether responses for the four scalar GCI types differed from those for contradictions and entailments. Recall that the FACT for the scalar implicature items was designed to contradict the implicature licensed by Sam’s utterance. So, if a scalar implicature item patterned like the contradiction control items – in which the FACT contradicts Sam’s utterance – then the implicature was taken to be incorporated into the truth-conditional meaning of the utterance. On the other hand, if a scalar implicature item patterned like the entailment control items – in which the FACT entails Sam’s utterance – then the implicature was taken to be distinct from truth-conditional meaning.

Our first model, Model 1, had item type (cardinals, quantificational items, ranked orderings, gradable adjectives, contradictions, entailments) as the only fixed effect and used entailments as the reference level, thereby comparing entailments to the four scale types and contradictions. Model 2 also had item type as the only fixed effect but used contradictions as the reference level, so it compared contradictions to the four scale types and entailments. These models showed that both contradictions and entailments significantly differed from all four scale types and from each other (p<0.001). The percentages of...
false responses for the four experimental scale types and the two types of control items are illustrated in Figure 1.

3.3. Effects of Scale Type and Discourse Condition

Our next set of models examined differences across scale types and discourse conditions and the interactions between them. After conducting a preliminary analysis of the data, we built a full model (Model 3) with scale type, discourse condition, and their interaction as fixed effects, using gradable adjectives as the scale type reference level and (N) as the discourse condition reference level. The full model was compared to three simpler models using likelihood ratio tests, which compare how well the models fit the data. These comparison models were: Model 4 – only scale type as a fixed effect, Model 5 – only discourse condition as a fixed effect, and Model 6 – scale type and discourse condition as fixed effects, with no interactions between them.⁹ Model 3, Models 4-6 had the same reference levels as Model 3.
with both discourse condition and scale type as well as their interactions as fixed effects, fits the data significantly better than the three comparison models. Model 3 shows a significant difference between the (N) and (M) discourse conditions and significant differences between the gradable adjective scale type and the cardinal, quantificational, and ranked ordering scale types. Finally, this model, taking gradable adjectives and (N) as the reference levels, shows significant interactions for (M) and the three other scale types. The percentage of “false” responses for the four scale types in each of the three discourse conditions are illustrated in Figure 2.

The significant differences we observed between scale types and discourse conditions, as well as the significant interactions found in the full model – coupled with the fact that this model significantly outperformed models without discourse condition, scale type, and their interactions as fixed

---

**Figure 2.** Barplot showing the percentage of “false” responses for the three discourse conditions ((M), (O), and (N)) in the four scale types (cardinals, quantificational items, ranked orderings, and gradable adjectives). A higher percentage of “false” responses indicates that the implicatures associated with a category were more often included in truth-conditional meaning.

---

10 Model 3 vs. Model 4: \( \chi^2 (8) = 18.616, p<0.05 \); Model 3 vs. Model 5: \( \chi^2 (9) = 106.56, p<0.001 \); Model 3 vs. Model 6: \( \chi^2 (6) = 15.512, p<0.05 \).

11 These results are reported in Table 3, found in Appendix C.
effects – led us to explore these effects more fully with a series of additional models. We first examined the discourse condition effect for each scale type individually. We built a pair of models for each of the four scale types, for a total of eight models, with discourse condition as the only fixed effect. The first model in each pair had (N) as the reference level and the second model had (M) as the reference level. The relevant results from all eight of these models are summarized in Table 4, found in Appendix C. Only gradable adjectives showed significant differences between the discourse conditions: The (M) discourse condition differed significantly from the (O) and (N) conditions (Reference N:M Estimate: -1.3842, Std. Error: 0.4209, z-score: -3.289, p<0.005; Reference M:O Estimate: 0.881, Std. Error: 0.4071, z-score: 2.165, p<0.05).

Our finding that the (M) condition differs from the other two conditions for gradable adjectives alone raises the question of whether gradable adjectives in general license scalar implicatures in the absence of multiple scalar values being evoked in the context. To address this question, we built a model that only compared entailments to gradable adjectives in the (N) discourse condition. If entailments and gradable adjectives in the (N) discourse condition patterned the same, it is possible that scalar implicatures are not generated for gradable adjectives when no scalar values are mentioned in the context. However, entailments were significantly different from gradable adjectives in the (N) discourse condition (Estimate: -0.9169, Std. Error: 0.4027, z-score: -2.277, p<0.05), suggesting that our gradable adjective stimuli in the (N) condition, at least sometimes, generated implicatures.

In order to determine the overall differences between the four scale types, we used a set of three models with different scale type reference levels. Each model had scale type, discourse condition, and their interaction as fixed effects. The relevant results from these models are summarized in Table 5, found in Appendix C. Most of the scale types were highly significantly different from one another (p<0.001); however, quantificational items were only marginally significantly different from cardinals and ranked orderings (p<0.1). Given that discourse condition had a significant effect for only one scale type (gradable adjectives), we conducted an additional analysis on scale types in the (M) condition – the condition that evoked the greatest number of scalar values and thus, presumably, rendered the relevant scale most salient in context. We built a model using only items in the (M) discourse condition. This model had only scale type as a fixed effect, with gradable adjectives as the reference level. Gradable adjectives were still significantly different from quantificational items and cardinals (p<0.01), while gradable adjectives and ranked orderings were not significantly different from each other (p>0.1). These results are summarized in Table 6, found in Appendix C.
4. Discussion

The results from this experiment reveal two interacting factors that affect the interpretation of scalar implicature. The first of these is the number of scalar values evoked in the discourse context, but this factor does not affect our four scale types uniformly. Recall that we manipulated the context by varying the number of scalar values evoked, ranging from none to one to multiple alternative values evoked (our (N), (O) and (M) conditions, respectively). However, varying the context in this way affected truth value judgements only for gradable adjective items.

For gradable adjectives in the (N) condition, where Irene’s question did not evoke multiple scalar values, participants’ responses were different from the pattern of responses found in the entailment controls, indicating that the participants were at least sometimes incorporating the implicature into the truth-conditional meaning. However, in the (M) condition, in which multiple scalar values were evoked in Irene’s question, the frequency with which the implicatures were incorporated increased significantly. This finding suggests that the interpretation of implicatures associated with gradable adjectives, in contrast to those associated with the other scale types, is influenced by the number of scalar values evoked in the preceding context.

To explain the more frequent incorporation in the (M) condition, we note that Sam’s response containing a weaker scalar value was used in contrast to a stronger, more informative alternate value already explicitly evoked in the context. This effect suggests that the interpretation of gradable adjectives shifts from being only weakly scalar in the absence of a contrast with other scalar values, to being more strongly scalar when the context favours a more informative interpretation, i.e. one in which stronger values do not hold. In the other discourse conditions, Sam’s utterance could more readily be interpreted as not making a claim that a stronger scalar value was false because there was no explicit contrast among alternate scalar values.

While discourse condition played a significant role in the interpretation of gradable adjectives, it did not affect the interpretation of the other scale types: the response patterns in the (N), (O), and (M) conditions for cardinals, quantification items, and ranked orderings were not significantly different from each other. That is, for these types, the number of scalar values evoked in Irene’s question did not affect the frequency of incorporation into truth-conditional meaning. Even in the (N) and (O) conditions, where Sam’s response did not select between explicitly evoked stronger and weaker scalar values, participants nonetheless judged Sam’s utterance similarly to when a contrast was established. The interpretation of gradable adjectives, on the other hand, was more sensitive to the explicit evocation of multiple scalar values.
The immunity of cardinal, quantificational, and ranked ordering scales to the effects of discourse condition suggests that alternate values on these scales enjoy a greater degree of salience – once the scale has been evoked – than those on gradable adjective scales. The greater salience of alternate values on these three scale types can be explained by considering that their values are more closely associated with the other values on the scale, such that mentioning the scale or a value on that scale is more likely to evoke alternate values. Such an association follows from the relative domain independence of the lexical items used to refer to scalar values. For example, the relation between *some* and *most* is essentially domain-independent; that is, this relationship holds whether applied to atoms, books, or zebras. The same holds for cardinals and ranked orderings. However, stronger values on gradable adjective scale are domain-restricted. For example, *hot* is a shared value across temperature scales but stronger values are lexicalised differently in different domains, e.g. *sweltering* (atmospheric temperature), *sizzling* (food), *scalding* (liquids), or *burning up* (humans with a fever). Thus, once the scale is evoked, the values on the cardinal, quantificational, and ranked ordering scales are inherently salient, resulting in no effect of discourse condition on the likelihood of incorporation into truth-conditional meaning.

The restriction of certain lexical items to particular domains, however, cannot fully account for the frequency with which an implicature is judged to impact truth-conditional meaning. To see why this is so, we must consider the second factor that affects the interpretation of scalar implicature, namely scale type. We found two main significant differences among scale types: ranked orderings were incorporated into truth-conditional meaning less often than cardinals, and gradable adjectives were incorporated less than all other scale types.

To explain the first of these findings, we note that an important difference between our ranked ordering and cardinal stimuli was that all of the ranked orderings were presented in modal contexts that strongly favoured the “at least” interpretation. For example, one of our ranked ordering stimuli included the following modal statement by Sam: “Juniors can register for the advanced seminar”, which licenses the inference that seniors can register for it as well. Given that all scalar values are associated with both “at least” and “exactly” interpretations, the relevant issue is which factors lead participants to favour one over the other. Under the “exactly” interpretation, which in our experiment corresponded to the incorporation of the implicature into truth-conditional meaning, higher scalar values are semantically excluded and therefore cannot constitute part of what is said. In non-modal contexts, on the other hand, ranked orderings strongly disfavour the “at least” interpretation and therefore
we did not use such contexts in our stimuli. For example, the values corresponding to the conventional sizes of mattresses (*king, queen, full, twin*) do not lend themselves readily to the “at least” interpretation in a non-modal context: e.g. when John says that he bought a twin size mattress, he asserts, in the absence of special circumstances, that the mattress he bought is not a king size. However, in a modal context, such as discussing which size mattress can fit into a certain room, both interpretations are available: e.g. if a queen mattress can fit in a certain room, a king mattress may be able to as well.

Another possible reason for the difference between ranked orderings and cardinals in their frequency of incorporation into truth-conditional meaning is the way the stimuli for the two scale types were constructed. The FACTs for the ranked ordering stimuli were written in such a way as to make reference to sets of values rather than to a specific higher scalar value. We did this because, as noted above, the “exactly” interpretation for ranked orderings is generally favoured and the FACT needed to capture all weaker scalar values to overcome this preference. For example, in (18) below, reference to a very large room in the FACT conveys that any size mattress will fit in the guest room.

(18) Irene: Will a twin, full, queen, or king mattress fit in the unfurnished guest room?
    Sam: A full mattress will fit in the guest room.
    FACT: The unfurnished guest room is as big as a two car garage.

In the context of (18), the FACT does not pick out any particular scalar value; rather, any scalar value on the relevant scale would be true. That is, what is relevant in our ranked ordering stimuli is that the entire set of values, rather than one specific scalar value, is inferrable from the FACT. In the cardinal stimulus provided in (19), on the other hand, the FACT does pick out one specific scalar value, i.e. four.

(19) Irene: Does Lisa have two, three, four, or more children?
    Sam: Lisa has three children.
    FACT: Lisa has quadruplets.

Thus, the relative infrequency of incorporation for ranked orderings, as compared to cardinals, can be further explained because the FACT picks out a set that includes the value in Sam’s utterance. The more inclusive FACT for this scale type encourages the “at least” interpretation of Sam's utterance, which resists the incorporation of the corresponding implicature.

Our second scale type finding was that implicatures associated with gradable adjectives are less frequently incorporated into truth-conditional meaning than all other scale types. Values on gradable adjective scales share with cardinal
and quantificational scales the property of entailing weaker values, e.g. something being hot entails that it is warm. However, they differ from cardinal and quantificational scales in different ways. On the one hand, gradable adjectives are unlike cardinals because gradable adjectives, like quantificational items, have vague boundaries. For example, the boundaries of hot on the temperature scale are much less clear than the boundaries of three on the cardinal scale. On the other hand, unlike quantificational scales, gradable adjective scales have no upper bound. For example, there is no maximally loud volume (on the gradable adjective scale of loudness), but no student can attend class more often than always (on the quantificational scale of frequency). The combination of vague boundaries and a lack of an upper bound permit lexically unmarked, non-maximal values to range indefinitely higher, without excluding stronger scalar values. This means that the values on gradable adjective scales tend to be interpreted non-exclusively, e.g. a hot day may be a felicitous description of a sweltering day, while Many people had a good time is generally an infelicitous description of everyone having a good time. Scalar values that are interpreted less exclusively are correspondingly less likely to favour the “exactly” interpretation. Therefore, scalar implicatures associated with gradable adjectives are less likely to be incorporated in the truth-conditional meaning of utterances containing these values.

One way to increase the exclusivity of scalar values for gradable adjectives is by evoking multiple scalar values in context by using them in contrast with one another, as in our (M) condition. Recall that implicatures associated with gradable adjectives in the (M) condition were significantly more likely to be incorporated into truth-conditional meaning than those in either the (O) or (N) condition. However, even in the (M) condition, when they were most likely to be considered part of truth-conditional meaning, participants nonetheless incorporated them significantly less often than cardinals and quantificational items, which had higher rates of “false” responses. This shows that even when used in a context that encourages an interpretation exclusive of stronger values, gradable adjective implicatures are not incorporated into truth-conditional meaning as often as the other implicature types.

These explanations of our findings assume, broadly speaking, that a greater frequency of incorporation into truth-conditional meaning corresponds to a more frequent generation of implicatures or to an increase in the strength of the implicatures generated. Under this assumption, a higher rate of “false” responses, when participants judged that the implicature impacted truth-conditional meaning, indicates that the implicature was generated more often or was stronger than in the cases where the percentage of “false” responses was lower. While it is clear that an implicature was generated in cases where it
was incorporated into truth-conditional meaning, this experimental paradigm cannot distinguish between cases in which an implicature was generated and subsequently cancelled, and cases in which an implicature was not generated at all. A further online experiment is needed to distinguish between these cases and to assess the extent to which discourse context and scale type impact the generation and subsequent interpretation of scalar implicature. The results from the current study do show, however, that the interpretation of scalar implicature is sensitive to both the associated scale type and discourse context.

Acknowledgements

The research presented in this paper was partially supported by a research grant from the University Research Grants Committee at Northwestern University. We would like to thank Matt Goldrick for his help with the statistical analyses, Matthew Berends and Alex Djalali for their help with the experiment, as well as audiences at the 2007 and 2008 LSA Annual Meetings, the 2007 Experimental Pragmatics Conference, the 2008 Invitational Symposium on Variation in English Conference, and the Northwestern University Philosophy and Linguistics Workgroup.

References


Jaeger, T. Florian. 2008. Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language* 59: 434-446.


Appendix A: Experimental Items

Cardinals

1) <…3, 2, 1>
   (N) Irene: How many books did Harry buy?
       Sam: He bought two books.
       FACT: Harry bought the Lord of the Rings trilogy.
   (M) Irene: Did Harry buy one, two, three books or more?
       Sam: He bought two books.
       FACT: Harry bought the Lord of the Rings trilogy.
   (O) Irene: Did Harry buy three books?
       Sam: He bought two books.
       FACT: Harry bought the Lord of the Rings trilogy.

2) <…6, 5, 4…>
   (N) Irene: How many roses do you need to make the bouquet?
       Sam: I need five of them.
       FACT: Sam needs a half-dozen roses to make the bouquet.
   (M) Irene: How many roses do you need to make the bouquet?
       Four, five, six, or more?
       Sam: I need five of them.
       FACT: Sam needs a half-dozen roses to make the bouquet.
   (O) Irene: Do you need six roses to make the bouquet?
       Sam: I need five of them.
       FACT: Sam needs a half-dozen roses to make the bouquet.

3) <…4, 3, 2…>
   (N) Irene: How many children does Lisa have?
       Sam: Lisa has three children.
       FACT: Lisa has quadruplets.
   (M) Irene: Does Lisa have two, three, four, or more children?
       Sam: Lisa has three children.
       FACT: Lisa has quadruplets.
(O) Irene: Does Lisa have four children?
   Sam: Lisa has three children.
   FACT: Lisa has quadruplets.

4) <…36, 24, 12…>

(N) Irene: How many doughnuts did the secretary order for the meeting?
   Sam: He ordered 24.
   FACT: The secretary ordered 3 dozen doughnuts.

(M) Irene: Did the secretary order 12, 24, 36, or more doughnuts for the meeting?
   Sam: He ordered 24.
   FACT: The secretary ordered 3 dozen doughnuts.

(O) Irene: Did the secretary order 36 doughnuts for the meeting?
   Sam: He ordered 24.
   FACT: The secretary ordered 3 dozen doughnuts.

5) <…20, 10, 5…>

(N) Irene: How many years have Edna and Tom been married?
   Sam: They’ve been married 10 years.
   FACT: Edna and Tom have been married for two decades.

(M) Irene: Have Edna and Tom been married 5, 10, 20 years or longer?
   Sam: They’ve been married 10 years.
   FACT: Edna and Tom have been married for two decades.

(O) Irene: Have Edna and Tom been married 20 years?
   Sam: They’ve been married 10 years.
   FACT: Edna and Tom have been married for two decades.

6) <…100, 90, 80…>

(N) Irene: How many points did the Bulls score in last night’s basketball game?
   Sam: The Bulls scored 90 points.
   FACT: The Bulls made exactly 35 2-pointers and 10 3-pointers in last night’s basketball game.

(M) Irene: Did the Bulls score 80, 90, 100, or more points in last night’s basketball game?
   Sam: The Bulls scored 90 points.
FACT: The Bulls made exactly 35 2-pointers and 10 3-pointers in last night’s basketball game.

(O) Irene: Did the Bulls score 100 points in last night’s basketball game?
    Sam: The Bulls scored 90 points.
    FACT: The Bulls made exactly 35 2-pointers and 10 3-pointers in last night’s basketball game.

7) <…20, 10, 5…>

(N) Irene: How many cigarettes does Abe smoke a day?
    Sam: He smokes 10 a day.
    FACT: Abe has smoked a pack of cigarettes every day for the past 18 years.

(M) Irene: Does Abe smoke 5, 10, 20, or more cigarettes a day?
    Sam: He smokes 10 a day.
    FACT: Abe has smoked a pack of cigarettes every day for the past 18 years.

(O) Irene: Does Abe smoke 20 cigarettes a day?
    Sam: He smokes 10 a day.
    FACT: Abe has smoked a pack of cigarettes every day for the past 18 years.

8) <… 10, 9, 8 …>

(N) Irene: How many cats does the crazy lady downstairs have?
    Sam: She has nine cats.
    FACT: The crazy lady that lives downstairs has five pairs of cats.

(M) Irene: Does the crazy lady downstairs have eight, nine, ten or more cats?
    Sam: She has nine cats.
    FACT: The crazy lady that lives downstairs has five pairs of cats.

(O) Irene: Does the crazy lady downstairs have ten cats?
    Sam: She has nine cats.
    FACT: The crazy lady that lives downstairs has five pairs of cats.

9) <…6, 5, 4…>

(N) Irene: How many bottles of beer did you bring over?
    Sam: I brought five.
    FACT: Sam brought a case of beer over.
(M) Irene: Did you bring over four, five, six, or more bottles of beer?
   Sam: I brought five.
   FACT: Sam brought a case of beer over.

(O) Irene: Did you bring six bottles of beer over?
   Sam: I brought five.
   FACT: Sam brought a case of beer over.

10) <…4, 3, 2…>

(N) Irene: How many car tires did your cousin sell on Craigslist?
   Sam: He sold two.
   FACT: Sam’s cousin sold a full set of car tires on Craigslist.

(M) Irene: Did your cousin sell two, three, four, or more car tires on Craigslist?
   Sam: He sold two.
   FACT: Sam’s cousin sold a full set of car tires on Craigslist.

(O) Irene: Did your cousin sell four car tires on Craigslist?
   Sam: He sold two.
   FACT: Sam’s cousin sold a full set of car tires on Craigslist.

Quantificational Items

1) <all, most, some>

(N) Irene: How much cake did Gus eat at his sister’s birthday party?
   Sam: He ate most of it.
   FACT: By himself, Gus ate his sister’s entire birthday cake.

(M) Irene: Did Gus eat some, most, or all of his sister’s birthday cake?
   Sam: He ate some of it.
   FACT: By himself, Gus ate his sister’s entire birthday cake.

(O) Irene: Did Gus eat all of his sister’s birthday cake?
   Sam: He ate most of it.
   FACT: By himself, Gus ate his sister’s entire birthday cake.

2) <everyone, many, a few>

(N) Irene: How many of her classmates did Esther invite to her party?
   Sam: She invited many of them.
   FACT: Esther invited her whole class to her party.
M) Irene: Did Esther invite a few of her classmates, many of her classmates, or everyone in her class to her party?
Sam: She invited many of them.
FACT: Esther invited her whole class to her party.

O) Irene: Did Esther invite everyone in her class to her party?
Sam: She invited many of them.
FACT: Esther invited her whole class to her party.

3) <always, frequently, sometimes>

N) Irene: How often does Jim go to class?
Sam: He frequently goes to class.
FACT: Jim has a perfect attendance record at school.

M) Irene: Does Jim sometimes, frequently, or always go to class?
Sam: He frequently goes to class.
FACT: Jim has a perfect attendance record at school.

O) Irene: Does Jim always go to class?
Sam: He frequently goes to class.
FACT: Jim has a perfect attendance record at school.

4) <definitely, probably, possibly>

N) Irene: How likely is Larry to come to your party?
Sam: He's probably coming to my party.
FACT: Larry has already RSVP’d to confirm he will attend Sam’s party.

M) Irene: So is Larry possibly, probably, or definitely coming to your party?
Sam: He's probably coming to my party.
FACT: Larry has already RSVP’d to confirm he will attend Sam’s party.

O) Irene: So is Larry definitely coming to your party?
Sam: He's probably coming to my party.
FACT: Larry has already RSVP’d to confirm he will attend Sam's party.

5) <permanently, a year, a month>

N) Irene: For how long is Tasha going to live in Morocco?
Sam: She's going for a year.
FACT: Tasha bought a house in Morocco and is going to live there forever.
(M) Irene: Is Tasha going to live in Morocco for a month, a year, or permanently?
    Sam: She's going for a year.
    FACT: Tasha bought a house in Morocco and is going to live there forever.

(O) Irene: Is Tasha going to live in Morocco permanently?
    Sam: She's going for a year.
    FACT: Tasha bought a house in Morocco and is going to live there forever.

6) <entire, a portion>

(N) Irene: How much of the wedding did Jody's parents pay for?
    Sam: They paid for a portion of it.
    FACT: Only Jody's parents paid for the wedding.

(M) Irene: Did Jody's parents pay for a portion or for the entire cost of the wedding?
    Sam: They paid for a portion of it.
    FACT: Only Jody's parents paid for the wedding.

(O) Irene: Did Jody's parents pay for the entire cost of the wedding?
    Sam: They paid for a portion of it.
    FACT: Only Jody's parents paid for the wedding.

Ranked Orderings

1) <senior, junior, sophomore, freshman>

(N) Irene: Undergraduates in which years can register for the advanced seminar?
    Sam: Juniors can register.
    FACT: Anyone who has completed his or her first year of studies can register for the advanced seminar.

(M) Irene: Can freshmen, sophomores, juniors, or seniors register for the advanced seminar?
    Sam: Juniors can register.
    FACT: Anyone who has completed his or her first year of studies can register for the advanced seminar.

(O) Irene: Can seniors register for the advanced seminar?
    Sam: Juniors can register.
FACT: Anyone who has completed his or her first year of studies can register for the advanced seminar.

2) <whole, 2%, 1%, skim> milk fat content

(N) Irene: What kind of milk does your weight loss diet allow for?
Sam: It allows for 1%.
FACT: The only type of milk prohibited by Sam's weight loss diet is full-fat milk.

(M) Irene: What kind of milk can you drink on your weight loss diet: skim, 1%, 2%, or whole?
Sam: It allows for 1%.
FACT: The only type of milk prohibited by Sam's weight loss diet is full-fat milk.

(O) Irene: Is 2% milk allowed on your weight loss diet?
Sam: It allows for 1%.
FACT: The only type of milk prohibited by Sam's weight loss diet is full-fat milk.

3) <doctoral, master's, bachelor's> level of education

(N) Irene: The new position requires a university degree. What level of education have you completed?
Sam: I've completed a master's level of education.
FACT: Sam recently successfully defended his dissertation on John Donne.

(M) Irene: The new position requires a university degree. Have you completed a bachelor's, master's, or doctoral level of education?
Sam: I've completed a master's level of education.
FACT: Sam recently successfully defended his dissertation on John Donne.

(O) Irene: The new position requires a university degree. Have you completed a doctoral level of education?
Sam: I've completed a master's level of education.
FACT: Sam recently successfully defended his dissertation on John Donne.

4) <king, queen, full, twin> mattress size

(N) Irene: What size mattress will fit in the unfurnished guest room?
Sam: A full mattress will fit in the guest room.
FACT: The unfurnished guest room is as big as a two car garage.

(M) Irene: Will a twin, full, queen, or king mattress fit in the unfurnished guest room?
Sam: A full mattress will fit in the guest room.
FACT: The unfurnished guest room is as big as a two car garage.

(O) Irene: Will a queen mattress fit in the unfurnished guest room?
Sam: A full mattress will fit in the guest room.
FACT: The unfurnished guest room is as big as a two car garage.

5) <advanced, intermediate, beginner> level of golf class

(N) Irene: I hear Erica is pretty good at golf. Which golf class is she qualified to teach?
Sam: She’s qualified to teach the intermediate level.
FACT: The Professional Golf Association has certified Erica to teach all levels of golf classes.

(M) Irene: I hear Erica is pretty good at golf. Is she qualified to teach the beginner, intermediate, or advanced level golf classes?
Sam: She’s qualified to teach the intermediate level.
FACT: The Professional Golf Association has certified Erica to teach all levels of golf classes.

(O) Irene: I hear Erica is pretty good at golf. Is she qualified to teach the advanced level golf class?
Sam: She’s qualified to teach the intermediate level.
FACT: The Professional Golf Association has certified Erica to teach all levels of golf classes.

**Graded Adjectives**

1) …gorgeous, pretty, average-looking…>

(N) Irene: How attractive is Kate?
Sam: She’s pretty.
FACT: Kate was voted World’s Most Beautiful Woman this year.

(M) Irene: Is Kate average-looking, pretty, or gorgeous?
Sam: She’s pretty.
FACT: Kate was voted World’s Most Beautiful Woman this year.
(O) Irene: Is Kate gorgeous?
   Sam: She's pretty.
   FACT: Kate was voted World's Most Beautiful Woman this year.

2) <…huge, big, average…>

(N) Irene: What size is Jeremy?
   Sam: He's big.
   FACT: Jeremy can't fit in an airplane seat.

(M) Irene: In terms of size, would you say that Jeremy is average, big, or huge?
   Sam: He's big.
   FACT: Jeremy can't fit in an airplane seat.

(O) Irene: In terms of size, would you say that Jeremy is huge?
   Sam: He's big.
   FACT: Jeremy can't fit in an airplane seat.

3) <…sweltering, hot, warm…>

(N) Irene: How was the weather in Barcelona yesterday?
   Sam: It was hot in Barcelona yesterday.
   FACT: The temperature was so high in Barcelona yesterday that it set new records, and the hospitals were inundated with people suffering from heat stroke.

(M) Irene: Was it warm, hot, or sweltering in Barcelona yesterday?
   Sam: It was hot in Barcelona yesterday.
   FACT: The temperature was so high in Barcelona yesterday that it set new records, and the hospitals were inundated with people suffering from heat stroke.

(O) Irene: Was it sweltering in Barcelona yesterday?
   Sam: It was hot in Barcelona yesterday.
   FACT: The temperature was so high in Barcelona yesterday that it set new records, and the hospitals were inundated with people suffering from heat stroke.

4) <wealthy, comfortable, getting by, poor>

(N) Irene: How would you say Alex is doing financially?
   Sam: He's comfortable.
   FACT: Alex just bought four condos at Lake Point Tower, in downtown Chicago, where Oprah Winfrey lives.
(M) Irene: Would you say that Alex is poor, comfortable, or wealthy?
   Sam: He's comfortable.
   FACT: Alex just bought four condos at Lake Point Tower, in downtown Chicago, where Oprah Winfrey lives.

(O) Irene: Would you say Alex is financially wealthy?
   Sam: He's comfortable.
   FACT: Alex just bought four condos at Lake Point Tower, in downtown Chicago, where Oprah Winfrey lives.

5) <…furious, upset, annoyed…>

(N) Irene: How does Roberta feel about her current situation with Joe?
   Sam: She's upset.
   FACT: Roberta is seething with resentment toward Joe and won't even talk to him.

(M) Irene: Is Roberta annoyed, upset, or furious with Joe?
   Sam: She's upset.
   FACT: Roberta is seething with resentment toward Joe and won't even talk to him.

(O) Irene: Is Roberta furious with Joe?
   Sam: She's upset.
   FACT: Roberta is seething with resentment toward Joe and won't even talk to him.

6) <…blaring, loud, audible…>

(N) Irene: How high is the music volume coming from the club next to your apartment on the weekends?
   Sam: It’s loud.
   FACT: The club next to Sam’s apartment gets cited by the police every weekend because the music volume is too high and causes severe noise pollution.

(M) Irene: Is the music from the club next to your apartment audible, loud, or blaring on the weekends?
   Sam: It’s loud.
   FACT: The club next to Sam's apartment gets cited by the police every weekend because the music volume is too high and causes severe noise pollution.

(O) Irene: Is the music from the club next to your apartment blaring?
   Sam: It’s loud.
FACT: The club next to Sam's apartment gets cited by the police every weekend because the music volume is too high and causes severe noise pollution.

7) <…bawling, teary, misty…>
   (N) Irene: I heard that book had a sad ending. How much did you cry at the end?
   Sam: I was teary.
   FACT: Sam cried so much when he read the end of the book that he drenched his shirt with tears.
   (M) Irene: I heard that book had a sad ending. Were you misty, teary or bawling at the end?
   Sam: I was teary.
   FACT: Sam cried so much when he read the end of the book that he drenched his shirt with tears.
   (O) Irene: I heard that book had a sad ending. Were you bawling at the end?
   Sam: I was teary.
   FACT: Sam cried so much when he read the end of the book that he drenched his shirt with tears.

8) <…ecstatic, happy, content…>
   (N) Irene: How good did you feel after passing the exam?
   Sam: I was happy.
   FACT: Sam jumped up and down and squealed with glee for 5 minutes after receiving his exam results.
   (M) Irene: Were you content, happy, or ecstatic to have passed the exam?
   Sam: I was happy.
   FACT: Sam jumped up and down and squealed with glee for 5 minutes after receiving his exam results.
   (O) Irene: Were you ecstatic to have passed the exam?
   Sam: I was happy.
   FACT: Sam jumped up and down and squealed with glee for 5 minutes after receiving his exam results.

9) <…brilliant, smart…>
   (N) Irene: How intelligent is your friend Ben?
   Sam: He's smart.
FACT: Ben has always been in gifted classes and has an IQ of 165.

(M) Irene: Is your friend Ben smart or brilliant?
Sam: He's smart.
FACT: Ben has always been in gifted classes and has an IQ of 165.

(O) Irene: Is your friend Ben brilliant?
Sam: He's smart.
FACT: Ben has always been in gifted classes and has an IQ of 165.

10) <… obnoxious, annoying, irksome…>

(N) Irene: How does your sister feel about her new roommate?
Sam: She thinks she's irksome.
FACT: Sam's sister finds her new roommate's personality so intolerable that she's looking for a new roommate.

(M) Irene: Does your sister think her new roommate is irksome, annoying, or obnoxious?
Sam: She thinks she's irksome.
FACT: Sam's sister finds her new roommate's personality so intolerable that she's looking for a new roommate.

(O) Irene: Does your sister think her new roommate is obnoxious?
Sam: She thinks she's irksome.
FACT: Sam's sister finds her new roommate's personality so intolerable that she's looking for a new roommate.

11) <…tight, snug…>

(N) Irene: How does your new bike helmet fit?
Sam: It's snug.
FACT: Sam's new bike helmet is two sizes too small for his head.

(M) Irene: Does your new bike helmet fit snugly or tightly?
Sam: It's snug.
FACT: Sam's new bike helmet is two sizes too small for his head.

(O) Irene: Does your new bike helmet fit tightly?
Sam: It's snug.
FACT: Sam's new bike helmet is two sizes too small for his head.
Appendix B: Control and Filler Items

Control items consisted of contradictions and entailments; filler items consisted of non-scalar GCIs (commitatives, conjunction buttressing, bridging inferences, and argument saturation) and pragmatically determined elements of what is said (deictics, ellipses, and indexicals). An example of each is provided below.

**Control Item: Entailment**

Irene: What was on the front page of today’s paper?
Sam: “Crystal Meth is a growing problem in Chicago”.
FACT: The headline on today’s newspaper reads “Crystal Meth is a growing problem in Chicago”.

**Control Item: Contradiction**

Irene: Tell me what Mark’s sister looks like, so I’ll recognize her when I pick her up at the station.
Sam: She’s a tall redhead.
FACT: Mark’s only sister is 4’7” with black hair.

**Filler Item: Commitative**

Irene: I heard something big happened in the art studio yesterday.
Sam: Yeah! In a fit of rage, Rachel picked up a hammer and broke a statue.
FACT: After grabbing a hammer, Rachel angrily kicked a statue, causing it to fall over and break.

**Filler Item: Bridging Inference**

Irene: How was the wedding?
Sam: The ceremony was nice. I spent the entire evening talking with the maid of honor.
FACT: Sam spent the evening talking to a woman who was the maid of honor in another wedding taking place at the same hotel.

**Filler Item: Conjunction Buttressing**

Irene: I understand that George has had a really rough year.
Sam: Yeah. Last month, he lost his job and started drinking.
FACT: George started drinking on the 15th of last month and lost his job on the 20th of last month.

**Filler Item: Deictic**

Irene: Have you seen my notebook?
Sam: It's over there.
FACT: Irene’s notebook is in Sam’s hand.

**Filler Item: Ellipsis**

Irene: What did the guys do yesterday?
Sam: Tom washed the SUV, and Joel did, too.
FACT: Tom washed the SUV alone, and Joel washed the motorcycle.

**Filler Item: Indexical**

Irene: Did you have your annual dentist appointment yet?
Sam: Yes. I went yesterday.
FACT: Sam went to his annual dentist appointment on the same day as this conversation.
### Appendix C: Statistical Analyses

#### Table 1. Parameter values for the fixed effect item type in a mixed logistic regression model of True/False responses, in log odds, and associated standard errors, z-scores, and probabilities. The scale types adjectives, cardinals, quantificational items, and ranked orderings, and the control type contradictions are compared to the reference level entailments.

<table>
<thead>
<tr>
<th>Item Type</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>z-score</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.8365</td>
<td>0.2766</td>
<td>10.254</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Adjectives</td>
<td>-1.2548</td>
<td>0.3680</td>
<td>-3.410</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Cardinals</td>
<td>-3.3227</td>
<td>0.3635</td>
<td>-9.141</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Quantificational Items</td>
<td>-2.8545</td>
<td>0.3635</td>
<td>-9.500</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Ranked Orderings</td>
<td>-1.7210</td>
<td>0.3934</td>
<td>-4.374</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Contradictions</td>
<td>-7.0160</td>
<td>0.3831</td>
<td>-18.312</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

##### Table 2. Parameter values for the fixed effect item type in a mixed logistic regression model of True/False responses, in log odds, and associated standard errors, z-scores, and probabilities. The scale types adjectives, cardinals, quantificational items, and ranked orderings, and the control type entailments are compared to the reference level contradictions.

<table>
<thead>
<tr>
<th>Item Type</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>z-score</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-4.1795</td>
<td>0.3674</td>
<td>-11.376</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Adjectives</td>
<td>5.7611</td>
<td>0.4403</td>
<td>13.085</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Cardinals</td>
<td>3.6932</td>
<td>0.4365</td>
<td>8.460</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Quantificational Items</td>
<td>4.1615</td>
<td>0.4509</td>
<td>9.229</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Ranked Orderings</td>
<td>5.2950</td>
<td>0.4618</td>
<td>11.466</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Entailments</td>
<td>7.0160</td>
<td>0.3831</td>
<td>18.312</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

##### Table 3. Parameter values for the fixed effects discourse condition, scale type, and their interaction in a mixed logistic regression model of True/False responses, in log odds, and associated standard errors, z-scores, and probabilities. The discourse conditions (O) and (M) are compared to the reference level (N), and the scale types cardinals, quantificational items, and ranked orderings are compared to the reference level adjectives.

<table>
<thead>
<tr>
<th>Item Type</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>z-score</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.2228</td>
<td>0.3409</td>
<td>6.521</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>O</td>
<td>-0.4661</td>
<td>0.3234</td>
<td>-1.441</td>
<td>0.149482</td>
</tr>
<tr>
<td>M</td>
<td>-1.2866</td>
<td>0.3138</td>
<td>-4.101</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Cardinals</td>
<td>-2.7137</td>
<td>0.3214</td>
<td>-8.445</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
Table 4. Parameter values for the fixed effect discourse condition in mixed logistic regression models of True/False responses, in log odds, and associated standard errors, z-scores, and probabilities. This table summarises the relevant comparisons from eight regression models. Two models were created for each scale type (adjectives, cardinals, quantificational items, and ranked orderings) to allow comparisons between the three discourse conditions ((M), (N), and (O)). For each scale type, one model had (N) as the reference level for the discourse condition factor, and the other model had (M) as the reference level.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Error</th>
<th>z-score</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantificational Items - Reference (N): (O)</td>
<td>-2.1127</td>
<td>0.3613</td>
<td>-5.847</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Ranked Orderings - Reference (N): (O)</td>
<td>-1.4118</td>
<td>0.3875</td>
<td>-3.643</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Interaction: O &amp; Cardinals</td>
<td>0.4094</td>
<td>0.4315</td>
<td>0.949</td>
<td>0.342820</td>
</tr>
<tr>
<td>Interaction: M &amp; Cardinals</td>
<td>1.3013</td>
<td>0.4270</td>
<td>3.047</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>Interaction: O &amp; Quantificational Items</td>
<td>0.1288</td>
<td>0.4945</td>
<td>0.260</td>
<td>0.794548</td>
</tr>
<tr>
<td>Interaction: M &amp; Quantificational Items</td>
<td>1.2207</td>
<td>0.4918</td>
<td>2.482</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Interaction: O &amp; Ranked Orderings</td>
<td>0.8859</td>
<td>0.5449</td>
<td>1.626</td>
<td>0.104005</td>
</tr>
<tr>
<td>Interaction: M &amp; Ranked Orderings</td>
<td>1.8411</td>
<td>0.5395</td>
<td>3.413</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Table 3. (Cont.,)
Table 4. (Cont.,)

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Error</th>
<th>z-score</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantificational Items only</td>
<td>-0.3252</td>
<td>0.3629</td>
<td>-0.8960</td>
<td>0.370</td>
</tr>
<tr>
<td>– Reference (N): (O)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quantificational Items only</td>
<td>-0.2535</td>
<td>0.3626</td>
<td>-0.6993</td>
<td>0.484</td>
</tr>
<tr>
<td>– Reference (N): (M)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quantificational Items only</td>
<td>-0.07165</td>
<td>0.36596</td>
<td>-0.1958</td>
<td>0.845</td>
</tr>
<tr>
<td>– Reference (M): (O)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ranked Orderings only</td>
<td>0.3977</td>
<td>0.4395</td>
<td>0.9049</td>
<td>0.3655</td>
</tr>
<tr>
<td>– Reference (N): (O)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ranked Orderings only</td>
<td>0.3548</td>
<td>0.4576</td>
<td>0.7753</td>
<td>0.4382</td>
</tr>
<tr>
<td>– Reference (N): (M)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ranked Orderings only</td>
<td>0.0430</td>
<td>0.4456</td>
<td>0.097</td>
<td>0.92312</td>
</tr>
<tr>
<td>– Reference (M): (O)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5. Parameter values for the fixed effect scale type in mixed logistic regression models of True/False responses, in log odds, and associated standard errors, z-scores, and probabilities. This table summarises the relevant comparisons from three regression models. One model had adjectives as the reference level for the scale type factor, one had cardinals as the reference level, and one had ranked orderings as the reference level.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Error</th>
<th>z-score</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference Adjectives: Cardinals</td>
<td>-2.7137</td>
<td>0.3214</td>
<td>-8.445</td>
<td>p&lt;0.001</td>
</tr>
<tr>
<td>Reference Adjectives: Quantifiers</td>
<td>-2.1127</td>
<td>0.3613</td>
<td>-5.847</td>
<td>p&lt;0.001</td>
</tr>
<tr>
<td>Reference Adjectives: Ranked Orderings</td>
<td>-1.4118</td>
<td>0.3875</td>
<td>-3.643</td>
<td>p&lt;0.001</td>
</tr>
<tr>
<td>Reference Cardinals: Quantificational Items</td>
<td>0.60109</td>
<td>0.33420</td>
<td>1.799</td>
<td>0.072084</td>
</tr>
<tr>
<td>Reference Cardinals: Ranked Orderings</td>
<td>1.30194</td>
<td>0.36598</td>
<td>3.557</td>
<td>p&lt;0.001</td>
</tr>
<tr>
<td>Reference Ranked Orderings: Quantificational Items</td>
<td>-0.7008</td>
<td>0.4040</td>
<td>-1.735</td>
<td>0.082766</td>
</tr>
</tbody>
</table>
Table 6. Parameter values for the fixed effect scale type in mixed logistic regression models of True/False responses, in log odds, and associated standard errors, z-scores, and probabilities. This table summarises the relevant comparisons from a regression model using data only from the (M) discourse condition. This model has adjectives as the reference level for the scale type factor.

<table>
<thead>
<tr>
<th>(M) only - Reference</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>z-score</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjectives: Cardinals</td>
<td>-1.4741</td>
<td>0.2892</td>
<td>-5.098</td>
<td>p&lt;0.001</td>
</tr>
<tr>
<td>Adjectives: Quantificational Items</td>
<td>-0.8838</td>
<td>0.3240</td>
<td>-2.728</td>
<td>p&lt;0.01</td>
</tr>
<tr>
<td>Adjectives: Ranked Orderings</td>
<td>0.4230</td>
<td>0.3635</td>
<td>1.164</td>
<td>0.24454</td>
</tr>
</tbody>
</table>