A Novel Experimental Paradigm for Distinguishing Between 'What is Said' and 'What is Implicated'

Ryan Doran, Gregory Ward, Meredith Larson, Yaron McNabb[†] and Rachel E. Baker[‡]

Northwestern University
[†]University of Chicago
[‡] EF Education First

Corresponding author:

Gregory Ward
Department of Linguistics
Northwestern University
Evanston, IL 60208

1-847-491-7020

gw@northwestern.edu

A Novel Experimental Paradigm for Distinguishing Between 'What is Said' and 'What is Implicated'

Abstract

That there is a theoretical distinction between context-dependent and context-independent aspects of utterance interpretation has become a standard assumption in current theories of meaning; however, how and where to draw this distinction has been the subject of considerable debate. In the current study, we investigate whether speakers can systematically distinguish between what is said and what is implicated (Grice 1967/89) using a novel truth-value judgment paradigm across a wide range of implicature types. We found that, by providing participants with a clear set of judgment criteria, including the adoption of an objective third-person perspective, we were able to enhance their ability to distinguish conversational implicature from truth-conditional meaning. In addition, we found that none of the implicature types we investigated was either consistently incorporated into or consistently excluded from what is said. Instead, our findings revealed considerable variation in frequency of incorporation across implicature types in ways that do not correspond straightforwardly to the various taxonomies of implicature types proposed in the literature.

Keywords: semantics/pragmatics boundary, generalized conversational implicature, Grice, scalar implicature, experimental pragmatics

Acknowledgments

[*] 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 and Matthew Berends for their help with the statistical analyses of the data. We would also like to thank Matthew Berends and Alex Djalali for their help with various aspects of the experiment proper. We would also like to acknowledge with gratitude numerous helpful

comments and suggestions that we received from audiences at the 2007 and 2008 LSA Annual Meetings, the 2007 Experimental Pragmatics Conference, the 2008 Invitational Symposium on Variation in English Conference, the Third Annual Midwest Workshop on Semantics, and the Northwestern University Philosophy and Linguistics Workgroup. Finally, we would like to acknowledge and express our gratitude for the many useful comments from two anonymous referees.

0. Introduction. How and where to draw the boundaries between context-dependent and context-independent aspects of utterance interpretation has recently been the subject of intense debate. The classic Gricean approach sharply distinguishes an utterance's semantics (what is said) from its pragmatics (what is implicated). More recently, alternative formulations of the semantic-pragmatic distinction have been proposed that call into question the existence of a distinct pre-pragmatic level of truth-conditional meaning, arguing instead that pragmatic enrichment can, and does, intrude upon truth-conditional meaning. In this paper, we describe a new experimental paradigm designed to investigate speakers' ability to systematically distinguish what is said from what is implicated in a controlled setting. Specifically, we were interested in how the diverse types of implicature identified in the literature pattern with respect to their frequency of incorporation by participants into truth-conditional meaning. Our paradigm was applied to this broad range of implicature types with the goal of identifying factors that affect the frequency of incorporation. By providing participants with a consistent set of judgment criteria, we hypothesized that we could reduce the frequency of implicature incorporation. The extent to which implicatures are not incorporated would provide evidence for a distinct level of meaning corresponding to Grice's notion of what is said.

1. Theoretical background. Grice (1967/89) introduced the notion of implicature to account for how speakers are able to pragmatically communicate more than what is strictly said by their words. In uttering a sentence, a speaker says what her words conventionally mean but can communicate more to the audience via implicature. The distinction here roughly corresponds to a distinction between semantics and pragmatics, where semantics is concerned with truth-conditional meaning (what is said) and pragmatics includes additional levels of meaning (what is

implicated), which are not part of the utterance's truth conditions. Consider the following classic examples of implicature, provided in (1):

(1)a. I broke a finger yesterday. [cf. Grice 1967/89, p. 38]

Implicature: The finger is not the speaker's.

b. You ate some of the cookies. [=Horn 1984, p. 14, ex. (2b)]

Implicature: You did not eat all of the cookies.

c. Bill caused the car to stop. [=Levinson 2000, p. 39, ex. (19b)]

Implicature: Bill did not stop the car in the stereotypical/normal way.

Following Grice, the semantics of (1a) determines what the speaker says (that she has broken some finger or other), yet it typically conveys to the hearer that the broken finger in question is the speaker's own. This additional pragmatically-communicated information is what is conversationally implicated on the basis of what the speaker said.

Grice – and indeed everyone working on this topic – acknowledged that the notion of what is said must include some amount of pragmatic specification in order to arrive at a truth-evaluable proposition. Examples of such specification include lexical disambiguation and reference resolution. In (1a), for example, given the presence of the two indexicals, the hearer cannot determine what is said until the identity of the speaker and the day of utterance are known. Such examples constitute pragmatic elements of what is said, as they are required for generating any truth-conditional meaning. Conversational implicatures, in contrast, are additional components of meaning that are not required for an utterance to express a truth-evaluable proposition. With this criterion in mind, we distinguish conversational implicatures from those pragmatic elements of meaning that are necessary to arrive at a truth-evaluable proposition by referring to the latter as necessary contextual elements of meaning (NCEs).[1]

Among pragmatic aspects of meaning that are not necessary for an utterance to express a truth-evaluable proposition, Grice drew distinctions within the class of implicature. In this paper, however, we will only be concerned with generalized conversational implicatures (GCIs). As an example of a GCI, consider (1a) when used by a speaker to implicate that the broken finger in question is her own. GCIs are distinct from particularized conversational implicatures (PCIs), in that GCIs arise under normal circumstances unless there is something in the context to prevent them (as Grice (1967/89:37) put it, 'in the absence of special circumstances'). All conversational implicatures for Grice arise from the adherence (or blatant non-adherence) to general conversational maxims, which speakers are mutually assumed to be observing when cooperatively engaged in talk exchanges. Conversational implicatures are thus calculable according to conversational maxims and are by definition cancellable. That is, since conversational implicatures are not part of truth-conditional meaning, a speaker can deny, or cancel, an implicature without denying the truth of what is said. In (1a), for example, the speaker could proceed to cancel the licensed implicature without contradiction by adding a phrase like but the finger wasn't my own. For all those working in the Gricean tradition, cancellability is an essential diagnostic for the presence of a conversational implicature.

The category of implicature that Grice called 'GCI' has received much discussion in the theoretical literature. Neo-Griceans, such as Horn (1984, 1993) and Levinson (2000), defend the classical Gricean view of GCIs in which they do not (by and large) constitute part of truth-conditional meaning and arise as unmarked or default meanings. Against this traditional understanding, Post-Griceans, such as Sperber & Wilson (1985), Récanti (1993), and Carston (2002), hold that some elements of meaning that are classified as GCIs by Grice can in fact 'intrude' upon truth-conditional meaning. Under this view, when extra-semantic aspects of

utterance interpretation do intrude upon the truth-conditions of an utterance, they become part of what is said.[2] Nonetheless, Post-Griceans retain cancellability as a crucial diagnostic for the identification of extra-semantic meaning. As with any pragmatically-determined aspect of utterance meaning, a speaker can deny such meaning without contradiction.[3]

2. Empirical investigations into GCIs. The investigation of GCIs has produced many studies that considered whether and under which conditions such implicatures arise in controlled experimental settings. The first study to directly address the question of whether speakers systematically distinguish between what is said and what is implicated is Gibbs & Moise 1997. This study describes several experiments that were designed to test, among other things: (1) whether participants are able to distinguish between 'minimal' (what is said) and 'enriched' (GCI-incorporated) interpretations and, if not, (2) whether participants, upon receiving explicit instruction, could learn to make Grice's original distinction between saying and implicating. In the first experiment, participants were asked to select between two paraphrases for each stimulus item – the minimal and the enriched interpretations – on the basis of which one 'best reflected what each sentence said' (Gibbs & Moise 1997:58). Participants' responses in this experiment overwhelmingly favored the enriched, GCI-inclusive interpretations as constituting what each sentence 'said'. The second experiment replicated the methodology of the first but with added instructions illustrating Grice's saying/implicating distinction by means of examples. Even with the additional instructions explaining Grice's technical use of 'saying', participants still overwhelmingly chose the enriched interpretations as best reflecting what they thought the stimuli 'said'. In subsequent experiments, however, participants were able to distinguish PCI

interpretations from what is 'said'. These results were taken to show that participants were not able to isolate a distinct level of meaning corresponding to Grice's notion of what is said.

Nicolle & Clark (1999) responded directly to Gibbs & Moise (1997) but were unable to replicate their main findings. In an effort to make methodological improvements, they used three conditions in which participants were asked to select paraphrases that best reflected 'what the speaker said', 'what the speaker's words meant', or 'what the speaker wanted to communicate'. Despite using the same types of GCIs as in the Gibbs & Moise study, Nicolle & Clark found that, in each of the three conditions, participants nonetheless favored the PCI interpretation over the enriched GCI-incorporated interpretation. Such results suggest that speakers do not routinely draw a distinction between saying and implicating, let alone distinguish between what is said and GCIs. Nicolle & Clark address the saying vs. implicating distinction in a subsequent experiment that included an important improvement to the methodology of Gibbs & Moise 1997: They had participants choose between interpretations based on the technical term 'saying', which was explained at the outset of the experiment. Gibbs & Moise (1997) also provided participants with an explanation of saying and implicating; however, their instructions for the experiment did not explicitly ask participants to use these terms to guide their responses, allowing them to employ their own notion of the term saying. The results of Nicolle & Clark's subsequent experiment showed that participants could distinguish between PCI-based interpretations and those based on GCIs. However, the paraphrases provided in this experiment did not include a 'minimal proposition' (i.e. one that corresponds to what is said); instead, all their paraphrases corresponded to either a PCI-based or GCI-based interpretation. Thus, this experiment does not provide data that address the question of whether speakers consider GCIs to constitute part of

what is said, although it does suggest that when given consistent criteria speakers can distinguish PCIs from GCIs.

Bezuidenhout & Cutting (2002) addressed participants' judgments concerning what is said by employing a methodology similar to Gibbs & Moise (1997) and Nicolle & Clark (1999). In the second of their four experiments, participants were asked to select the response that best reflected 'what was said', 'meant by the words', or 'communicated', but they were not given criteria to guide their judgments. Participants chose among paraphrases that included, among other things, a PCI-inclusive interpretation, a GCI-inclusive interpretation, and a 'minimal paraphrase', which was intended to correspond to Grice's what is said. The findings of this experiment were consistent with the findings of Nicolle & Clark's first experiment, as the PCI-inclusive paraphrase was selected most often. The findings were also consistent with Gibbs & Moise's findings since the 'minimal paraphrase' was the least favored. While this study benefits from the methodological improvement of offering participants what is said paraphrases among the possible choices, it is unclear what ultimately guided their judgments. As was the case with Nicolle & Clark 1999, these findings were not significantly different across instruction conditions, suggesting that whatever criterion guided participant responses it is not one that corresponds to a distinction between 'said', 'meant', and 'communicated'.

The authors of these studies have generally interpreted their findings as evidence for the claim that speakers do not systematically distinguish between the Gricean notion of what is said and GCIs. However, there are a number of problems with drawing such a conclusion based on the methodologies employed in these studies. First, either participants in these studies were not asked to use the saying/implicating distinction to guide their responses, or they were asked but were not provided with a choice between Grice's what is said and GCI-inclusive interpretations.

Thus, the question of whether GCIs are incorporated into what is said was not specifically addressed – even in those experiments that provided criteria to guide participants' judgments. Second, the experimental materials used were limited to a narrow range of GCI types and failed to include the wide variety of types that have been standardly classified as GCIs in the literature. Moreover, the types of GCI stimuli used in these experiments were not representative of the well-established taxonomies offered by Neo-Gricean approaches; rather, they were predominantly based on Grice's (1967/89) second Maxim of Quantity (classified as I-based implicatures by Levinson (2000) or R-based implicatures by Horn (1984)). Subsequent studies concerning the processing of GCIs (to be discussed below) have, in contrast, focused almost exclusively on scalar implicatures derived from Grice's first Maxim of Quantity (or Q-based implicatures in the taxonomies of Levinson (2000) and Horn (1984)).[4] Third, these studies focused on speakers' judgments concerning the meaning of 'what is said' and similar phrases, rather than the theoretical notions that these terms were employed to denote. It is not clear, however, whether the phrase 'what is said' as used in these experiments corresponds to the theoretical notion 'what is said' as used in the literature to identify a component of utterance meaning.

More recently, various empirical investigations of GCIs have explored a number of related questions regarding the psychological reality and availability of GCIs, including the acquisition of GCI-based meanings by children and the time-course of GCI processing. In the developmental literature on GCIs, children were found to be more 'logical' than adults when interpreting utterances that licensed scalar implicatures (Noveck 2001; Papafragou & Musolino 2003; Papafragou & Tantalou 2004; Guasti et al. 2005; Pouscoulous 2007, inter alia). That is, children were less likely to assign an upper-bounded interpretation (in the sense of Horn 1984) to

certain scalar terms than their adult counterparts.[5] For example, in a context in which it is true that all of a given set of horses jumped over a fence, children were more likely than adults to accept the sentence Some of the horses jumped over the fence as being true. As with the earlier studies discussed above, most of the acquisition studies of GCIs have employed a narrow range of GCI types, focussing on scalar terms.[6]

Other studies have focused on the processing of GCIs in adults by investigating response latency in categorization tasks. In Noveck & Posada 2003 and Bott & Noveck 2004, for example, participants differed as to whether they responded 'logically' or 'pragmatically' in judging whether 'underinformative' sentences such as Some cats are mammals were true. In these studies, a response of 'true' (i.e. the 'logical' response) indicates that the GCI was not incorporated into what is said despite its underinformativeness. A response of 'false' (i.e. the 'pragmatic' response), on the other hand, indicates that the GCI was incorporated into what is said (i.e. 'Some – but not all – cats are mammals'). Participants who responded 'logically' took less time to do so than those who responded 'pragmatically', from which the authors of these studies conclude that the incorporation of GCIs involves additional processing costs. This result is consistent with the findings of Chevallier et al. 2008 in which participants had limited time (i.e. one second) to respond and were found to respond more 'logically' than when required to wait at least three seconds before responding. Similarly, De Neys & Schaeken (2007) found that when participants were required to perform a second task simultaneously they were more likely to favor the 'logical' reading.

Further evidence that the processing of the 'pragmatic' (GCI-inclusive) interpretation is delayed relative to the processing of the 'logical' interpretation is provided by Huang & Snedecker (2009, 2011), who employ an eye-tracking methodology in a reference resolution

task. They found that when the identification of the target referent required an 'upper bounded' (GCI-inclusive) interpretation, participants were slower to identify the referent than when the task did not require such an interpretation. For Huang & Snedecker, the additional processing time required indicates that the 'pragmatic' interpretation of the quantifier some (i.e. 'some but not all') is subsequent to the initial semantic processing of the quantifier and that, therefore, the GCI associated with some constitutes an additional, extra-semantic inference.

Taken together, the results of the on-line studies discussed above show that the interpretation of GCIs involves additional processing costs. A theoretical consequence of these findings is that GCIs constitute a distinct type of meaning – one that is accessed subsequent to (and not parallel with) the lower-bounded readings of certain scalar terms. Although the evidence suggests that speakers distinguish GCIs from other types of meaning, the precise relationship between GCIs and truth-conditional meaning remains unclear. Specifically, it is an open question whether the results reported in these studies for the scalar quantifier some generalize to other types of GCIs and to what extent speakers can exclude the various GCI types from what is said.

3. A new experimental paradigm. In the previous section, we identified various methodological issues with previous off-line studies that addressed the saying/implicating distinction. We also highlighted some results from a number of on-line studies that have bearing on this distinction. In this section, we present a new experimental paradigm designed to address the methodological and theoretical issues raised in the previous section.

Our main methodological concern with previous studies pertains to the use of technical terms in the instructions given to participants. Even in studies in which participants were familiarized with the technical use of terms such as what is said, they were not encouraged to

base their judgments on the technical sense of those terms, and no particular interpretation strategy was mandated. Rather than allowing participants to devise their own interpretive strategy, we designed our paradigm in such a way as to give participants clear and consistent criteria to guide their interpretive strategy.

Having addressed this methodological issue, we can now consider our central theoretical question: whether speakers can systematically – and without reliance on technical terminology – distinguish a level of meaning corresponding to the Gricean notion of what is said exclusive of GCIs. Previous studies have produced mixed results on this score, with some suggesting that speakers do not, with others suggesting that they do for the narrow range of GCI types that were employed. To investigate the potential incorporation of GCIs into what is said, we asked participants to assess the truth of a statement with respect to a particular state of affairs that was compatible with what was said but incompatible with the corresponding GCI. In this way, instead of asking our participants whether a given GCI is included in what is said, we investigated the extent to which that GCI could be excluded from it. If participants' assessments of the truth conditions exclude the associated GCI, it will provide evidence for their ability to distinguish GCIs from what is said.

On the assumption that speakers do systematically distinguish between GCIs and what is said, a further theoretical question is whether they also distinguish among the various GCI types associated with different conversational maxims. Because previous studies included only a small number of different GCI types, it is unclear if the results of these studies generalize to all GCI types. To address this question, we specifically included a comprehensive array of GCI types in our materials, as will be explained in §4.

The paradigm we propose here for exploring the relationship between GCIs and what is said uses a truth-value judgment task that does not conflate technical and non-technical uses of terminology and that could be applied across the full spectrum of GCI types. The paradigm was designed to have participants rely on the folk notion of 'literal interpretation' in assessing the truth of various statements that could or could not be interpreted as GCI-inclusive. In addition, to keep participants' perspectives constant, we introduced a character named Literal Lucy from whose perspective participants in one condition were instructed to evaluate the truth of utterances. In what follows, we first discuss the specific design changes we made to previous methodologies and then present the specific questions that guided our study.

3.1 Experimental design. Our goal was to test whether speakers – with minimal training – can isolate a level of utterance interpretation independent of GCIs. To answer this question, we designed our experiment to focus participants' attention on truth-conditional meaning by suspending their natural tendency to supplement what is said in order to arrive at the speaker's intended meaning. To facilitate the task, we asked participants to rely upon the notion of truth rather than ask them to interpret what an utterance 'said', 'meant', or 'communicated'. In this way, we thought participants would be more likely to ignore speakers' intentions and focus only on what is required to make a given utterance true. The truth-value judgment task allowed for greater experimental control in terms of the context, interpretation of the target utterance, and the mapping of the utterance to the state of affairs (Crain & Thornton 1998). This task has been an effective paradigm for investigating putative implicatures and has been used in many of the previously mentioned studies concerned with the processing of implicature (Noveck & Posada 2003, Bott & Noveck 2004, inter alia) and the acquisition of implicature by children (Papafragou & Musolino 2003, Guasti et al. 2005, inter alia).

Taking the truth-value judgment task as our starting point, we developed a paradigm in which we attempted to standardize participants' behavior through minimal training and task instructions. We hypothesized that two factors would facilitate participants' use of a consistent interpretive strategy: (i) having them employ the folk-notion of interpreting something literally and (ii) shifting participants' perspective. This second factor stems from the observation (Dias et al. 2005) that speakers, both children and adults, tend to interpret utterances more 'logically' when they suspend their assumptions based on world knowledge or previous experience. Dias et al. (2005) asked adults to reason about utterances based on knowledge either outside of the adult's domain or in conflict with what the adult knows. All other things being equal, when adults reason about propositions that run counter to their experience, they tend to respond on the basis of their experiences rather than on logical entailments. For example, Dias et al. (2005) found that if adults were presented with the story in (2), they were likely to respond 'yes'.

(2) All blood is blue. John cut himself and bled on his shirt. Was his blood red?

However, if the adults were first told to shift their perspective and evaluate the story as though it were happening in another world, they were more likely to respond logically and answer 'no'. This suggests that a shift in perspective allows adults to attend more to the logical features of the discourse, even when the specified situation runs counter to their own experience.

In our study, we elected to integrate the use of perspective-taking with the folk notion of interpreting something literally to create three different instruction conditions (discussed in greater detail in §4.2). To shift participants' perspectives, we used a character named Literal Lucy from whose perspective participants were instructed to evaluate the truth of various utterances. We predicted that by evoking participants' notion of interpreting literally and by using a third-person perspective, we would create an interpretation strategy which participants

could use to evaluate the truth of utterances consistently. This paradigm was designed to provide an experimental setting that was maximally conducive to participants' being able to distinguish GCIs from what is said.

Our experiment also included an internal measure to provide feedback about the task itself by employing a confidence rating. It has been shown that participants' reported degree of confidence in their response in the performance of some task often correlates (inversely) with task complexity. Fellbaum et al. (1998), for example, provided participants with a list of possible semantic categories with which to annotate words, and when the choices among the categories were more similar (making the choice harder) participants rated their confidence lower. The authors argue that the confidence ratings correlate with both the actual and perceived difficulty of the task. Thus, ratings from this measure will allow us to gauge the level of difficulty associated with our materials and the different instruction conditions.

3.2 Questions guiding the current study. While the primary motivation for this study was an empirical investigation of the class of GCIs and the distinction between what is said and what is implicated, two specific questions guided our study: (1) to what extent do the instructions that participants receive affect their truth-value judgments?; and (2) given appropriate instructions, (a) do participants systematically exclude GCIs from truth-conditional meaning?; and (b) do participants' responses distinguish among the various GCI types discussed in the literature? In the following sections, we address each of these questions in turn.

4. Overview of the experiment. Our empirical investigation of GCIs was designed to provide participants with a set of tasks that would measure their ability to isolate a level of meaning corresponding to what is said without their reliance upon theory-dependent concepts. For each

stimulus item, we had participants first read a short conversation, then make a truth-value judgment about a target sentence in that conversation based on a particular set of instructions, and finally rate their confidence in their judgment. All of the participants read the same conversations, which were divided into four stimulus types. However, each participant was presented with only one of three sets of instructions. Thus, there was one within-participants factor with four levels (i.e. the four stimulus types) and one between-participants factor with three levels (i.e. the three different instructions for the truth-value judgment task). We will first discuss the stimulus types in §4.1, followed by a discussion of the instruction conditions in §4.2.

4.1 Stimuli. The stimuli consisted of 11 types of GCIs and 2 control types (Entailments and Contradictions). In addition to the GCI and control types, we also included four types of necessary contextual elements (NCEs). Each of these types was presented in a conversational format that consisted of an exchange between two characters, Irene and Sam. In the course of each conversation, Irene asked Sam a question to which he responded. The target portion of Sam's response, indicated by underlining, was always the final sentence of the conversation. Underneath the conversation was a sentence labelled 'FACT', providing a context for the target sentence.

Participants first saw the entire conversation between Irene and Sam along with the FACT on a computer monitor (see Stage #1 in Figure 1 below). They were instructed to press the space bar after having completed reading the conversation at their own pace. Next, they were asked whether the underlined sentence was true given the FACT, with the conversation remaining visible (see Stage #2 in Figure 1).[7] Once participants responded by pressing either 'T' ('true') or 'F' ('false') on the keyboard, a 4-point Likert scale appeared underneath (see Stage #3 in Figure 1). After participants assigned a confidence rating using a scale ranging from

1 to 4, with 1 signifying ‘not at all confident’ and 4 signifying ‘completely confident’, the screen would go blank, and the next conversation would appear.

INSERT FIGURE 1 ABOUT HERE

The experiment consisted of 88 conversations, including 28 control items (14 Entailments and 14 Contradictions), 44 GCIs (4 tokens of 11 types), and 16 NCEs (4 tokens of 4 types). The GCI items were adapted from examples of the various types of GCI discussed in the literature. The Entailment and Contradiction control items are discussed in §4.1.1 below, while the GCI and NCE items are discussed in §4.1.2 and 4.1.3, respectively.[8]

4.1.1 Entailments and Contradictions The Entailment and Contradiction control items were sentences whose truth values were not in question in light of the FACT presented in the conversation. As such, they served three useful purposes. First, they allowed us to determine whether participants were attending to the truth-value judgment task. Second, they offered a safeguard against response bias, as participants were presented both with items that clearly ought to elicit a ‘true’ response as well as items that clearly ought to elicit a ‘false’ one. Third, they provided a baseline to which the experimental items can be compared, thus providing a basis for determining whether particular GCIs are included in truth-conditional meaning.

For the Entailment items, illustrated in (3) below, the underlined sentence is entailed by the FACT and, therefore, should be judged by participants to be true.

(3) <u>Entailment item</u>

Irene: I haven’t seen Ed in ages. Have you?

Sam: No. <u>Ed lives in California.</u>

FACT: Ed lives in Los Angeles.

An example of a Contradiction item is provided in (4).

(4) Contradiction item

Irene: When did Robert's great-uncle Jake die?

Sam: He died in 1963.

FACT: Robert's great-uncle Jake died in 1957.

For the Contradiction items, the underlined sentence is contradicted by the FACT and, therefore, should be judged by participants to be false.

4.1.2 GCIs. The 11 types of GCIs used in our study were intended to be representative of the various types of GCIs discussed in the literature. Previous studies (e.g. Horn 1984 and Levinson 2000) have proposed different classification systems for GCIs based on reanalyses of Grice's original four conversational maxims.[9] Here, we adopt the taxonomy proposed by Levinson (2000), given his extensive discussion and cataloguing of the various GCI types. Levinson's classification system is based on three inferential heuristics – Q, I, and M – each of which gives rise to a distinct class of implicature, as discussed in turn below.

Q-based GCIs. Levinson's class of Q-based implicatures consists of scalar implicatures, which arise from Grice's first Maxim of Quantity and are licensed when a speaker uses a non-maximal value on some salient scale to convey that stronger values are either false or unknown (Horn 1984, Hirschberg 1991, inter alia). The Q-based implicature types we used as stimuli included both Horn scales (entailment-based scales) and Hirschberg scales (non-entailment-based scales), namely, quantifiers and modals, cardinals, gradable adjectives, and rankings. Examples of each of our four types of Q-based implicatures are provided in (5).

(5) Q-based implicatures

a. 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.

b. Cardinals

Irene: How many children does Lisa have?

Sam: Lisa has three children.

FACT: Lisa has quadruplets.

c. Gradable adjectives

Irene: How was the weather in Barcelona yesterday?

Sam: It was hot in Barcelona yesterday.

FACT: The temperature was so high in Barcelona that it set new records, and the hospitals were flooded by people suffering from heat stroke.

d. Rankings

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.

For all Q-based stimuli, the question posed by Irene was designed to evoke the relevant scale without explicitly mentioning any specific value on that scale. In (5a), for example, Irene's question (How much cake did Gus eat at his sister's birthday party?) evokes the quantifier scale (<all, most, some>) without specifically mentioning any value on it. Similarly, the

corresponding FACT never makes direct reference to a value on the scale, employing instead a phrase that is truth-conditionally equivalent to asserting a higher scalar value. For example, in (5b), the FACT (Lisa has quadruplets) entails the scalar value four without explicitly mentioning it. In this way, scalar values only appeared in the target sentence, as the mention of alternate scalar values have been shown to affect participant responses (Doran et al. 2009).

I-based GCIs. Levinson's I-based implicatures arise from Grice's second Maxim of Quantity and are based on the assumption that the speaker has said only what is necessary, leaving the hearer to infer a more specific and informative interpretation than that provided by the sentence's literal meaning alone. We used a total of four different I-based implicatures: argument saturation, bridging inferences, co-activities, and conjunction buttressing. Examples of each are provided in (6).

(6) I-based implicatures

a. Argument saturation

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.

b. Bridging inferences

Irene: What happened when Sue came over?

Sam: She walked into the bathroom. The window was open.

FACT: The open windows are in the kitchen, and there are no windows in the bathroom.

c. Co-activities

Irene: Can the guys come to the reception?

Sam: No. George and Steve play squash at the gym until 6:00 everyday.

FACT: George plays squash at the YMCA until 6:00 daily, and Steve plays squash at SPAC until 6:00 everyday.

d. 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.

In each of these examples, Sam's utterance licenses an I-based implicature that, according to Levinson, results in a more specific and informative interpretation. For example, the most natural interpretation of the underlined portion of Sam's utterance in (6a) is the more specific I-based interpretation that Rachel used the hammer in question to break the statue.

M-based GCIs. Levinson's M-based implicatures, derived from Grice's Maxim of Manner, are generated when a speaker selects a marked way of describing a certain state of affairs, thereby licensing the hearer to infer that the unmarked description does not apply and that, therefore, some feature of the state of affairs associated with the unmarked form does not hold. These implicatures arise from the hearer's consideration of alternative forms that the speaker might have employed, but did not. We used three different types of M-based GCIs, namely verbal periphrasis, repeated verb conjuncts, and repeated noun conjuncts; examples of each are provided in (7) below:

(7) M-based implicatures

a. Verbal periphrasis

Irene: How did Lynn avoid an accident when the deer darted into the road?

Sam: She made the car come to a halt.

FACT: Lynn slammed on the brakes.

b. Repeated verb conjuncts

Irene: What happened at Doctor Witherspoon's office?

Sam: Sasha waited and waited for her appointment.

FACT: Sasha waited 5 minutes for her appointment at Doctor Witherspoon's office.

c. Repeated noun conjuncts

Irene: What did Joseph do after finishing the marathon?

Sam: He drank bottles and bottles of water.

FACT: Joseph drank one 20 oz bottle and one 16 oz bottle of water after finishing the marathon.

In these examples, the speaker's choice of a marked expression (e.g. waited and waited in (7b)) over an unmarked one (e.g. waited) licenses the inference that the unmarked expression is inappropriate and that the speaker is, therefore, describing a non-stereotypical state of affairs (i.e. that the waiting time in question was greater than expected).

4.1.3 Necessary Contextual Elements. As discussed in §1, Grice (1967/89) acknowledged that certain aspects of what is said are pragmatically determined – that is, just those aspects of meaning that are required to arrive at a truth-evaluable proposition. We have been referring to such contextually-determined aspects of what is said as necessary contextual elements (NCEs). Of theoretical relevance for the current study is whether the interpretation of NCEs patterns like that of GCIs, given that they are both contextually-determined aspects of meaning, with the

former being uniformly part of what is said and the latter being only variably so. The question then arises: Can participants in a controlled experiment be directed to interpret NCEs in a way that is strongly dispreferred in context? As far as we know, the status of NCEs as a necessary component of propositional meaning has not been systematically investigated empirically. To compare the interpretation of NCEs and GCIs with respect to what is said, we included 4 different types of NCE phenomena among the experimental stimuli: deixis, ellipsis, indexicality, and pronoun resolution. An example of each is provided in (8).

(8) Necessary Contextual Elements (NCEs)

a. Deixis

Irene: What shoes are you wearing to dinner?

Sam: I'm going to wear these shoes.

FACT: Sam has decided to wear the shoes in the upstairs closet, not the ones he is currently putting on.

b. Ellipsis

Irene: What did everyone eat?

Sam: Robert ate apples, oranges, and pears, and so did Melissa.

FACT: Robert ate apples, oranges, and pears, and Melissa ate only apples and oranges.

c. Indexicality

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.

d. Pronoun resolution

Irene: I haven't seen that coat I gave you for Christmas... And what did you do with the sweater I gave you?

Sam: I hung it in the closet.

FACT: Sam hung the coat Irene gave him in the closet, and he put the sweater from Irene in his dresser drawer.

All NCE stimuli were designed so that the FACT contradicts the preferred interpretation of Sam's utterance containing the NCE. For example, in (8c), if Sam went to his annual dentist appointment on the day of the conversation, then the preferred interpretation of his utterance is false.

4.2 Experimental instructions and the truth-value judgment task. As mentioned above in §4, participants were assigned to one of three different instruction conditions, making instructions a between-participants factor. Each instruction condition was accompanied by a different training procedure. The instructions differed in the way they directed participants to evaluate the truth of the target sentence. We first introduce the Baseline instruction condition followed by the other two instruction conditions.

4.2.1 Baseline condition. In this condition, participants received training prior to the experiment proper in which they were simply told to evaluate the truth of target (underlined) sentence. As part of their training, they were first provided with the example in (9).

(9) Training example

Irene: Hey, Sam. Do you know who wrote *Pride and Prejudice*?

Sam: A British woman wrote it, and her last name was Austen.

FACT: Jane Austen, a British woman, wrote *Pride and Prejudice*.

Participants were then told that the information presented in the FACT was true and that Irene and Sam may or may not be aware of this FACT. Specifically, they were told that the underlined sentence was true 'since the FACT states that Jane Austen is a British woman who wrote "Pride and Prejudice" '. During the actual experiment, participants in this condition saw a conversation accompanied by a FACT and were then presented with the following prompt: 'Given this FACT, the underlined sentence is: T or F'. After completing the truth-value judgment task, participants rated their certainty using a Likert Scale, as described above.

4.2.2. Literal condition. The second instruction condition utilized the same pre-experiment training and examples as in the Baseline condition above with two key differences. First, the word literally was used in the instructions, and second, the wording for the truth-value judgment task prompt was also changed to include the word literally. Specifically, participants in this condition were told that the underlined sentence was true 'since the FACT literally states that Jane Austen is a British woman who wrote "Pride and Prejudice" '. Following the presentation of the conversation and FACT, they were given the prompt 'Given this FACT, the underlined sentence when interpreted literally is: T or F.'. As in the Baseline condition, participants in this condition were instructed to rate their confidence in their judgments on a 4-point Likert Scale. Our inclusion of the word literally in the training and the instructions was intended to direct participants to draw upon their pre-theoretical folk notion of interpreting literally without having to rely on technical terminology.

4.2.3. Literal Lucy condition. The third instruction condition also differed from the others with respect to the training given to participants and in the wording of the truth-value judgment task prompt. The training in this condition introduced participants to Literal Lucy, a fictional character whose literal-mindedness leads her to misinterpret instances of non-literal language,

such as figurative expressions and indirect speech acts. Participants read a number of examples, illustrated in (10), in which Literal Lucy demonstrates her propensity to interpret only the conventional meaning of the words uttered.

(10) Training example for the Literal Lucy condition

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?

Here, instead of interpreting the idiom to be over the hill idiomatically (i.e. 'to be past one's prime'), Literal Lucy interprets it literally (i.e. 'to be over a contextually unique hill'). After demonstrating how Literal Lucy interprets utterances, we then instructed participants to indicate how Literal Lucy would respond when evaluating the truth of the target sentences. The truth-value judgment task prompt for this condition was: 'Given this FACT, Literal Lucy would say that the underlined sentence is: T or F.'. Once again, participants then rated their confidence in their responses on a 4-point Likert Scale. As in the Literal condition, our use of the folk notion of interpreting literally was intended to guide participants to draw upon their pre-theoretical folk notion without having to rely on technical terminology. However, in the Literal Lucy condition, we asked participants to predict how someone else – in this case a literal-minded individual – would evaluate the truth of the target sentences. Our goal in doing so was to provide consistent, participant-external criteria by which to evaluate the truth of the target sentences, rather than allow them to rely on their own – possibly idiosyncratic – criteria for interpreting the speaker's intended meaning.

4.3 Block design and presentation. We used a block design for the experiment for two reasons: (1) to distribute the Contradiction and Entailment items so that there was some guarantee of variation in 'true' and 'false' responses and (2) to minimize the probability that participants would encounter a series of highly similar target items in a row.[10] The ordering of the conversations was randomized within blocks, and the presentation of blocks was counterbalanced such that by the midpoint of the experiment (i.e. after the presentation of two blocks) participants had encountered equal numbers of Entailments and Contradictions. Participants were tested individually in a sound-attenuated computer booth. Participants progressed through the experiment at their own pace, reading each conversation and FACT on the screen before moving to the truth-evaluation task and the Likert scale.

4.4 Participants. A total of 74 native speakers of North American English from the Northwestern University community participated for pay or course credit. Two participants were excluded from the analysis, as they diverged by two standard deviations from the means for responses to the Entailment and Contradictions items. Recall that we used performance on these items as a diagnostic for whether participants understood and were performing the truth-value judgment task correctly. Thus, data from 72 participants (24 per condition) were analyzed.

5. Results and discussion. The discussion of the results proceeds as follows. First, we present the confidence ratings and their implications for our design. Then we present the truth-value data as they pertain to the questions mentioned in §3.2.

5.1 Results from Likert scale ratings. Recall that we are using the Likert data as a measure of task difficulty, where lower confidence scores are assumed to correlate with greater perceived effort (Fellbaum et al. 1998). In this way, the confidence scores indicate the difficulty as judged

by participants with any of the instruction conditions or stimulus types. We would expect to find relatively high confidence ratings (and, therefore, less effort) associated with the Contradiction and Entailment control items given that the evaluation of their truth values does not require additional extra-semantic inferencing, while the additional inferencing associated with GCIs would lead us to expect lower confidence ratings for these items. Similarly, we would expect relatively high confidence ratings for the Baseline instruction condition, with progressively declining confidence ratings as the instruction conditions include additional criteria – that is, the change to 'literally true' in the Literal condition and the introduction of a third person perspective in the Literal Lucy condition. Of particular interest to us is whether we would find significant differences between participants' confidence ratings for the GCI items across the instruction conditions.

5.1.1 Likert scale: Results. Confidence ratings were made using a 4-point Likert scale, with 1 indicating 'not at all confident' and 4 indicating 'completely confident'. We submitted the ratings to a logistic regression with stepwise (forward) coding of the instruction conditions and dummy coding of the stimuli with GCIs as the baseline. The participants and stimuli were random effects, and the instruction conditions and stimulus types were fixed effects in an interaction model.[11] By using stepwise coding, we were able to compare the Baseline condition to the Literal condition and then the Literal condition to the Literal Lucy condition. Thus, we could determine whether, first, the addition of interpreting literally and, then, the addition of interpreting literally from another (literally minded) person's perspective showed stepwise increase in difficulty.

To determine whether there was a main effect for instruction condition, we collapsed the 4 stimulus types (i.e. Entailments, Contradictions, GCIs, and NCEs) and found no significant

difference between the Baseline and the Literal conditions (estimate = 0.03, p = 0.56, CI [-0.07, 0.13]) or between the Literal and the Literal Lucy conditions (estimate = 0.06, p = 0.31, CI [-0.05, -0.16]).[12] Overall, participants in the Literal Lucy condition were less confident in their ratings (mean = 3.28) than those in the Literal (mean = 3.47) or the Baseline condition (mean = 3.49); however, this difference did not reach significance.

As for the main effect of stimulus type, we found that the GCI stimuli led to significantly lower confidence ratings (mean = 3.19) than did each of the other three stimulus types: NCEs (mean = 3.41; estimate = 0.22, $p < 0.001$, CI [0.11, 0.33]), Entailments (mean = 3.72; estimate = 0.53, $p < 0.001$, CI [0.41, 0.65]), and Contradictions (mean = 3.83, estimate = 0.63, $p < 0.001$, CI [0.51, 0.76]).[13] The difference in confidence ratings between NCEs and Entailments also reached significance (estimate = 0.31, $p < 0.001$, CI [0.17, 0.46]), as did that between NCEs and Contradictions (estimate = 0.42, $p < 0.001$, CI [0.28, 0.56]).

Although there was no main effect of instruction condition, we did find significant differences between the instruction conditions for Contradictions and NCEs, as illustrated in Figure 2.

INSERT FIGURE 2 ABOUT HERE

For Contradictions, there was a significant difference between the Baseline and Literal instruction conditions (estimate = 0.63, $p < 0.001$, CI [0.51, 0.76]), as well as between the Literal and Literal Lucy instruction conditions (estimate = 0.12, $p < 0.001$, CI [0.06, 0.20]). Likewise, for NCEs, there was a significant difference between the Baseline and Literal instruction conditions (estimate = 0.15, $p < 0.001$, CI [0.08, 0.22]), as well as between the Literal and Literal

Lucy instruction conditions (estimate = 0.26, p < 0.001, CI [0.20, 0.33]). As for GCIs and Entailments, however, we found no significant difference across instruction conditions.

5.1.2 Likert scale: Discussion. As discussed above in §3.1, we are interpreting the confidence ratings as indicators of perceived difficulty with instruction condition or stimulus type. In light of the results reported above, we can conclude that participants perceived GCI items to be more difficult than the control items, as predicted. We interpret this difficulty to be the result of participants' considering whether GCIs are incorporated into truth-conditional meaning. However, based on the lack of significant differences found among the instruction conditions, we conclude that the addition of either 'literally true' or the Literal Lucy character to the Baseline instruction condition did not, in and of itself, render the task more difficult for the participants, a finding that we did not predict. As for significant differences between instruction conditions within each stimulus type, we found none for the GCI items, although there were significant differences between instruction conditions for the NCE and Contradiction items. From this, we conclude that for GCIs as a whole the differences found in response patterns to the truth-value judgment task are not due to task difficulty. One explanation for the absence of significant differences between instruction conditions for the GCI items might be that the participants' perceived difficulty interpreting GCIs is already high in comparison to other stimulus types, and the additional complexity associated with the Literal Lucy condition doesn't further impact the perceived difficulty of the task.

5.2 Results from truth-value judgment task. Having established that the GCI stimuli were not significantly more effortful for our participants in the Literal Lucy instruction condition than in the other two instruction conditions, we now turn to a discussion of the truth-value judgment task itself. The purpose of this task was to investigate the extent to which participants can

exclude GCIs from what is said. In what follows, we present the data in terms of the percentage of 'false' responses provided by participants. A 'false' response to an item indicates a participant's judgment of incompatibility between the truth of the underlined statement and the FACT. For the Entailment control items, there should be a very low rate of 'false' responses because the FACT entails the truth of Sam's statement. On the other hand, for the Contradiction control items, there should be a very high rate of 'false' responses because the FACT contradicts the underlined statement. Thus, we expect our control items to serve as endpoints with respect to participants' responses: Contradictions are predicted to be at ceiling with respect to their rate of 'false' responses and Entailments to be at floor.

For the GCI experimental items, a 'false' response indicates that the participant judged that the FACT contradicted the GCI-inclusive interpretation of the underlined statement and that, therefore, the GCI was incorporated into truth-conditional meaning. A 'true' response, on the other hand, indicates that the participant judged the FACT to be compatible with the truth of the underlined sentence, indicating that a GCI-exclusive interpretation was assigned to the underlined sentence.[14] For example, consider again (5a), repeated below for convenience:

(11) Irene: How much of the 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. [=5a]

Sam's utterance, with the scalar NP most of it, has both a GCI-inclusive and a GCI-exclusive interpretation. If the GCI is incorporated into the truth-conditional meaning, then the underlined sentence is true only if Gus ate most – but not all – of the cake. Conversely, if the GCI is not incorporated into the truth-conditional meaning, then the underlined sentence is still true even if Gus all of the cake. For all GCIs, the FACT entails the GCI-exclusive interpretation but

contradicts the GCI-inclusive interpretation. Therefore, when participants respond 'false', we take this to indicate that they have incorporated the GCI into truth-conditional meaning.

We predicted that the particular task instructions presented to participants would affect their ability to isolate a level of meaning corresponding to what is said. Specifically, we predicted that the use of 'literally true' would enhance this ability and that the introduction of a literal-minded third person in the instructions would further enhance this ability. Thus, we hypothesized that the participants in the Literal condition would provide fewer 'false' responses to GCIs than those in the Baseline condition would. Furthermore, we hypothesized that participants in the Literal Lucy condition would provide even fewer 'false' responses to GCI. This pattern of predicted responses is illustrated in (12):

(12) Prediction of 'false' response rate for GCIs across the 3 instruction conditions:

Baseline > Literal > Literal Lucy

Once we identify the instruction condition that best facilitates participants' ability to isolate GCI-exclusive interpretations, we will then be in a position to make comparisons among the various stimulus types.

5.2.1 Instruction condition: Results. The results from the truth-value judgment task address the questions raised in §3.2. The first question concerned the extent to which task instructions affect participants' truth-value judgments, especially with respect to GCIs.

To explore this question, we analysed the GCI data using a logistic regression with participants and stimuli as the random factors and instruction condition as the fixed factor. We chose the GCI data specifically to test the prediction illustrated in (12). We used the same stepwise (forward) coding of the instruction conditions as was used in the Likert data analysis. By using stepwise coding, we were again able to first compare the Baseline condition to the

Literal condition and then the Literal condition to the Literal Lucy condition in order to determine the possible additive effects of introducing 'literally true' and Literal Lucy to the instructions.

The regression analysis revealed an effect of condition at each step of the comparison. The response rate for the Literal condition was significantly different from that of the Baseline condition ($z = 2.35$, $p < 0.02$) and, further, the response rate for the Literal Lucy condition was significantly different from that of the Literal condition ($z = 2.69$, $p < 0.01$). Participants in the Baseline condition were more likely to evaluate Sam's statement as false than those in the Literal condition (50% vs. 44%). Similarly, participants in the Literal condition were more likely to response 'false' than those in the Literal Lucy condition (44% vs. 36%). From these results, we see that participants in the Literal Lucy condition were least likely to incorporate GCIs into truth-conditional meaning, as illustrated in Figure 3.

INSERT FIGURE 3 ABOUT HERE

The low rate of implicature incorporation associated with the Literal Lucy condition indicates that this is the instruction condition that best facilitates participants' ability to exclude GCIs from truth-conditional meaning. Henceforth, we focus our analysis on data from that condition.

5.2.2 Stimulus type: Results. Having established that instruction type had a significant effect on participants' responses and that the Literal Lucy instruction condition was the most likely to facilitate their ability to exclude GCIs from truth-conditional meaning, we are now in a position to address our second question. That is, whether participants can systematically distinguish

between GCIs and truth-conditional meaning and, furthermore, whether their GCI response patterns suggest a distinction among the various GCI types as discussed in the literature.

In order to test whether participants can distinguish between GCIs and other stimulus types, we analyzed participants' responses from the Literal Lucy instruction condition using a generalized linear mixed model logistic regression with participants and stimuli as random factors and stimulus type as the fixed factor. We used dummy coding of stimulus type, with GCIs as the baseline. GCIs were significantly different from all of the other stimulus types, having a significantly higher 'false' rate than Entailments (36% versus 7%, $z = 6.07$, $p < 0.001$), a significantly lower 'false' rate than NCEs (36% versus 86%, $z = 9.06$, $p < 0.001$), and a significantly lower 'false' rate than Contradictions (36% versus 99%, $z = 7.95$, $p < 0.001$).

INSERT FIGURE 4 ABOUT HERE

As can be seen from Figure 4, GCIs as a whole patterned differently from the other stimulus types. Interestingly, NCEs also showed a distinct pattern with respect to the other stimulus types. An additional logistic regression with NCEs as the baseline (with stimulus type as a fixed effect and participants and stimuli as random effects) found that they are significantly different from Contradictions ($z = 3.57$, $p < 0.001$).

5.2.3 GCI type: Results. Although the finding that GCIs pattern significantly differently from the other stimulus types is noteworthy, of greater interest is whether GCIs form a coherent, uniform class or whether there is variation within it. Furthermore, if there is variation within the class of GCIs, can this variation be explained by the Maxims upon which the GCIs are based? Recall that in creating our stimuli, we relied on Levinson's (2000) taxonomy of Q-based

implicatures (related to Grice's first Maxim of Quantity), I-based implicatures (related to Grice's second Maxim of Quantity), and M-based implicatures (related to Grice's Maxim of Manner) in order to ensure a balanced variety of GCI types. A visual inspection of the response patterns for the various GCI types illustrated in Figure 5 reveals that participants' responses do not conform to a classification system based on Grice's conversational maxims. Rather, we see that the response rates for GCIs form a wide-ranging continuum ranging from frequent (63%) to infrequent (15%) incorporation.

INSERT FIGURE 5 ABOUT HERE

Figure 5 contains the average scores for each of the 11 GCI types and shows considerable variation within each GCI category.[15] For example, Q-based GCIs (gradable adjectives, quantifiers/modals, rankings, and cardinals) appear distributed across the continuum, including the items with the lowest average percentage of 'false' responses (gradable adjectives, 17%) as well as the items with the third highest average percentage of 'false' responses (cardinals, 53%). So we see that some Q-based GCIs, such as cardinals, are often incorporated into truth-conditional meaning (as compared to the overall mean for GCIs), while other Q-based GCIs, such as gradable adjectives, are only rarely so. Likewise, M-based GCIs seem to form two distinct groups based on the frequency with which they are incorporated into truth-conditional meaning. Repeated verb and noun conjuncts, for example, impact truth conditions for most participants (63% and 61% 'false' responses, respectively), while verbal periphrases do so for far fewer (24% 'false' responses). Finally, note that I-based GCIs are interspersed throughout the continuum of response rates. We discuss the implications of these finding in Section §6 below.[16]

5.2.4 Truth-value judgment task: Discussion. The data suggest that GCIs constitute a distinct class of meaning in that their response rates do not pattern like those of Entailments, Contradictions, or NCEs. By not patterning like the Entailments items, the response rates for GCIs suggest that GCIs are sometimes being incorporated into what is said. While at the same time, by not patterning like the Contradiction items, the response rates for GCIs suggest that it is not the case that GCIs are always incorporated into what is said. Finally, by not patterning like NCEs, the response rates for GCIs suggest that their interpretation is distinct from that of contextually-determined aspects of what is said. GCIs interpretation are not necessary for participants to evaluate the truth of a proposition.[17]

Taken together, our data suggest that participants are able to distinguish a level of truth-conditional meaning exclusive of GCIs. The ability to distinguish such meaning appears to be enhanced by both the use of participants' folk notion of interpreting literally and by the use of a third-person perspective. However, the continuum of 'false' responses within the GCI class suggests that there is no correlation between a particular conversational maxim from which a GCI is derived and the frequency with which it is incorporated into truth-conditional meaning in our truth-value judgment task. This can be seen quite clearly in the case of Q-based, or scalar, implicatures. For example, cardinals are likely to be incorporated whereas gradable adjectives are not. This suggests that variation among the scalar implicatures is a property of some other feature, perhaps properties of the scale itself, as argued by Doran et al. (2009).

6. General discussion. Our study was designed to investigate the relationship between GCIs and what is said in a controlled experimental setting. We were specifically interested in the extent to which the instructions that participants received affected their truth-value judgments.

Moreover, we were interested in determining whether participants, with appropriate instructions, could systematically distinguish between GCIs and truth-conditional meaning and whether this ability was sensitive to the various types of GCIs discussed in the literature. To address these issues, we developed a new paradigm in which participants were provided with clear and consistent criteria to guide their responses in a truth-value judgment task.

The results of our study show that GCI items were interpreted significantly differently from both the Contradiction and the Entailment control items. With respect to truth-conditional meaning, we found that no GCI type was consistently incorporated into what is said. By the same token, we also found that no GCI type was consistently excluded from what is said. Thus, we can conclude that for each GCI type, participants only sometimes incorporated the corresponding implicature into what is said, suggesting that speakers are able to access an interpretation exclusive of GCIs – that is, one in which the GCI does not affect the truth-conditional meaning of the utterance. This conclusion is consistent with the Neo-Gricean position that GCIs are additional pragmatic inferences, calculated on the basis of an utterance's truth-conditional meaning (i.e. Grice's what is said).[18] On the other hand, our study also reveals that, for some GCI types, participants routinely incorporate the relevant GCI into truth conditions even when instructed to interpret utterances literally. This finding is problematic for the Neo-Gricean position, as speakers do not consistently distinguish between GCI-inclusive and what is said interpretations. As for the Post-Gricean position, it cannot be further evaluated, at least with respect to our findings, without an explanation, presumably relevance-based, of the conditions under which participants incorporate, or fail to incorporate, GCIs into truth-conditional meaning.

The results of our study also show that participants' ability to isolate what is said was sensitive to instruction condition and type of GCI stimulus. In each of our instruction conditions,

participants succeeded in isolating what is said (albeit with varying frequency); however, the condition in which participants were directed to draw upon their pre-theoretical folk notion of interpreting literally (the Literal condition) enhanced their success as compared to the Baseline condition. In addition, the condition that prompted participants to adopt a literal-minded third-person perspective (the Literal Lucy condition) further enhanced their success in isolating what is said. We maintain that the inclusion of this perspective allowed participants to attend more to the strictly semantic interpretation of the discourse, even when that interpretation ran counter to the preferred one. The confidence ratings for the Literal Lucy condition reveal that this condition did not lead to greater perceived task difficulty for the participants with respect to GCIs. Thus, we conclude that the introduction of a third-person perspective (as in the Literal Lucy condition) enhanced participants' ability to isolate what is said (more than either of the other two instruction conditions), while at the same time not resulting in the task being perceived as more difficult.

As for the role of the various GCI types, we found that, depending on the particular type of GCI, participants displayed a considerable range in the frequency with which they incorporated GCIs into what is said, with mean rates of incorporation varying along a continuum ranging from 63% to 15%. The fact that the different GCI types form a continuum has both methodological and theoretical consequences. Methodologically, our findings suggest that researchers investigating GCIs need to include a broad range of GCI types if their goal is to generalize across GCI types lest their findings be unrepresentative of the phenomenon as a whole. On a more theoretical note, the taxonomies proposed in the literature based upon Gricean maxims do not correspond to the continuum of incorporation rates that we found in our study, nor do they explain the considerable variation we found within each GCI type (although n.b.

footnote #9). Given that the conversational maxims themselves do not provide an account of these findings, one must look elsewhere for an explanation.[19]

For Q-based implicatures, recall that we found a rate of incorporation that ranged from a low of 17% (for scalar adjectives) to a high of 53% (for cardinals). Thus, the frequency with which incorporation occurs within the class of Q-based implicatures cannot be explained by virtue of its associated maxim; we need to look elsewhere for explanations to account for the patterns we observed in our study. Of significance to the rate of incorporation may be some of the semantic and pragmatic features of the scales themselves. For example, Doran et al. (2009) found significant differences between cardinals and scalar adjectives with respect to the frequency with which their associated implicatures were incorporated into truth-conditional meaning. As they note, the boundaries of scalar adjectives are inherently vague as compared to the boundaries of cardinals. In this way, the inherent vagueness of scalar adjectives allows for the use of a weaker value to describe a state of affairs that could also be described by a stronger value on that scale. Thus, one can felicitously refer to a scorching day as 'hot', whereas one cannot felicitously say the temperature is 80 when it is fact 100. Moreover, Doran et al. note that for scalar adjectives, unlike cardinals, stronger values are lexicalized differently in different domains. For example, stronger scalar alternatives to 'hot' may be represented by 'sweltering' in the context of atmospheric temperatures or by 'scalding' in the context of liquid temperatures, but typically not vice versa. In contrast, the inherent precision and domain independence of cardinals leads to the upper-bounded ('exactly') interpretation being strongly favored across contexts and may account for the more frequent incorporation we found in our study. Indeed, the unique status of cardinals (either for adults or children, depending on the particular study) has been widely noted in other studies as well (e.g. Horn 1992; Carston 2002; Chierchia 2004;

Huang, Spelke, & Snedeker 2004, 2010; inter alia). No doubt other factors also play a significant role in affecting the conditions under which – and the frequency with which – scalar implicatures are incorporated into what is said.

To the best of our knowledge, no empirical study of implicature has included examples derived from the maxim of Manner. In our study, we used the three types of M-based implicatures – verbal periphrasis, repeated nouns conjuncts, and repeated verb conjuncts – each of which involved the use of a marked expression that licensed an inference to the effect that a non-stereotypical state of affairs obtains. Although our study included examples from only three types of M-based implicatures, our findings suggest that there is nonetheless a great deal of variation in how manner implicatures are interpreted with respect to truth conditions. For example, participants were much less likely to judge the verbal periphrasis Lynn made the car come to a halt to be false when told that Lynn slammed on the brakes, than they were to judge the repeated verb conjunct Sasha waited and waited to be false when told that Sasha waited only a short period of time (24% vs. 63% 'false' responses for verbal periphrasis and repeated verb conjuncts, respectively). One possible explanation for this finding is that the interpretation of M-based implicatures is sensitive to the degree of lexicalization for particular expressions and constructions. That is, the use of the expression wait and wait is conventionally associated with a considerable waiting time; indeed, the V_x and V_x construction generally conveys a large degree or amount. Examples of verbal periphrasis, on the other hand, such as make the car come to a halt, are not lexicalized ways of describing a particular state of affairs. Thus, our participants were more likely to respond 'false' in connection with stimuli like waited and waited because there is a typical state of affairs associated with the description and the FACT contradicted it. As a consequence of such an explanation, it is worth considering whether certain M-based

implicatures, namely those based on repeated noun and verb conjuncts, are generated by the same pragmatic processes as are other GCIs, as opposed to their being closely associated with the conventional meanings of the constructions.

Finally, as for I-based implicatures, associated with Grice's second maxim of Quantity, we found that they did not display as much variation as did either Q-based or M-based implicatures. Specifically, the range between the most frequently incorporated and the least frequently incorporated I-based implicature types was much smaller (only 24%) than the range found for Q- and M-based implicature types (with a 36% and 39% range between the most and least frequently incorporated types, respectively). Within I-based implicatures, two types – argument saturations and bridging inferences – were more likely to be judged false by participants (30% and 39%, respectively) than were the two other types – co-activities and conjunction buttressing (15% and 21%, respectively) – although our study does not have enough power to make statistically valid claims regarding significant differences between one implicature type and another. One possibility may be that the narrower range of incorporation rates that we found for I-based implicatures is simply an artefact of the specific examples that were used. An alternative, and more interesting, explanation is that the two types of I-based implicatures with lower rates of incorporation both crucially involve the conjunction and. Under this view, the presence of the conjunction licenses two easily identifiable interpretations: the logical (corresponding to what is said) and the enriched (corresponding to the GCI-incorporated interpretation). When focussing on what is strictly required for the truth of the target utterance, participants were less able to ignore the logical meaning of and, which, in turn, rendered the incorporation of the GCI into truth conditions correspondingly difficult. Further research will no

doubt uncover other factors that, more generally, affect the interpretation of each particular category of GCI and, more specifically, affect the incorporation of GCIs into what is said.

Another topic for future research is the status of NCEs. In the current study, we found that when participants were provided with additional criteria to consider when making truth-value judgments, their confidence ratings for the Contradiction and NCE items were lowered significantly as opposed to the Entailment and GCI items (whose confidence ratings were unaffected by the presence of additional criteria). In other words, confidence ratings for Contradictions and NCEs were significantly higher in the Baseline condition than they were in the Literal condition. Likewise, confidence ratings for Contradictions and NCEs were significantly higher in the Literal condition than in the Literal Lucy condition. One possible explanation for this is that, in general, participants sought reinterpretations for the target sentence of a Contradiction or NCE item rather than simply judging it to be false. With each additional evaluation criterion, participants reported lower confidence ratings for these items because they undertook greater effort in trying to reinterpret the target sentence. When asked if the target sentence was literally true, as in the Literal condition, participants were less confident that they had in fact interpreted the sentence literally. Furthermore, when adopting the perspective of Literal Lucy, participants were guided by the additional criterion of whether a literal-minded person would accept the target sentence as true; thus, additional processing effort was required to search for possible reinterpretations not required in the Baseline or Literal conditions.

However, despite the Literal Lucy condition being significantly more effortful for both NCEs and Contradictions, the results from the truth value judgment task show that participants' response patterns for these stimulus types differed significantly from one another, with Contradictions at ceiling but NCE stimuli not judged false in all cases. Our findings show that –

in at least some cases – participants succeeded in reinterpreting the target NCE items so that the underlined sentence would be true, but they were generally unable to do so for Contradictions. Thus, this suggests that in some circumstances speakers can be induced to assign an NCE a possible, yet highly implausible, interpretation. Such findings suggest additional avenues of research into the conditions under which speakers are likely to seek strongly dispreferred interpretations.

An additional point worth making is that NCEs were grouped together on the basis of what is required for there to be a truth-evaluable proposition, but these elements may too form a heterogeneous class, as was seen in the case of GCIs. For example, there is reason to think that ellipsis is syntactically (or semantically) controlled; therefore, non-preferred interpretations for ellipsis may be much less available (if at all) than non-preferred interpretations for, say, discourse pronouns. Another interesting question is whether seemingly related phenomena like indexicality, deixis, and discourse anaphora show differences with respect to the conditions under which, or the ease with which, speakers seek to reinterpret their truth-conditional contribution. Although theorists have rightfully claimed that what we are calling NCEs are required for <u>what is said</u>, our results suggest that further empirical study of contextual, yet truth-conditionally obligatory, aspects of meaning is called for as well.

A final direction for further study concerns the paradigm employed in the current study. Our results show that participants treated GCIs as a distinct class of meaning and we intend to augment the current study with a less strategic paradigm (e.g. measuring readings times, eye-tracking studies) in order to see whether the findings from more automatic paradigms are consistent with the findings reported here. Ideally, data from a broad range of paradigms will

converge and ultimately provide us with a clear and consistent set of criteria for empirically distinguishing the said from the implicated.

References

Bach, Kent. 1994. Conversational impliciture. Mind and Language 9(2):124-162.

Barner, David, Neon Brooks, and Alan Bale. 2011. Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. Cognition 118(1):84–93.

Bezuidenhout, Anne L. and Cooper J. Cutting. 2002. Literal meaning, minimal propositions, and pragmatic processing. Journal of Pragmatics 34(4):433-456.

Bezuidenhout, Anne L. and Robin K. Morris. 2004. Implicature, relevance, and default pragmatic inference. Experimental Pragmatics, ed. by Dan Sperber and Ira Noveck. Hampshire: Palgrave MacMillan, 257-282.

Bott, Lewis and Ira A. Noveck. 2004. Some utterances are underinformative: The onset and time course of scalar inferences. Journal of Memory and Language 51(3):437-457.

Breheny, Richard, Napoleon Katsos, and John Williams. 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.

Carston, Robyn. 2002. Thoughts and Utterances: The Pragmatics of Explicit Communication. Malden, MA: Blackwell.

Chevallier, Coralie, Ira A. Noveck, Tatjana Nazir, Lewis Bott, Valentina Lanzetti, and Dan Sperber. 2008. Making disjunctions exclusive. Quarterly Journal of Experimental Psychology 61(11):1741–1760.

Chierchia, Gennaro. 2004. Scalar implicatures, polarity, and the syntax/pragmatics interface. Structure and Beyond, ed. by Andriana Belletti. Oxford: Oxford University Press, 39-103.

Crain, Stephen and Rosalind Thorton. 1998. Investigations in Universal Grammar. Cambridge, MA: MIT Press.

De Neys, Wim and Walter Schaeken. 2007. When people are more logical under cognitive load: Dual task impact on scalar implicature. Experimental Psychology 54(2):128-133.

Dias, Maria, Antonio Roazzi, and Paul L. Harris. 2005. Reasoning from unfamiliar premises: A study with unschooled adults. Psychological Science 16(7):550-554.

Doran, Ryan, Rachel E. Baker, Yaron McNabb, Meredith Larson, and Gregory Ward. 2009. On the non-unified nature of scalar implicature: An empirical investigation. International Review of Pragmatics 1:211-248.

Fellbaum, Christiane, Joachim Grabowski, and Shari Landes. 1998. Performance and confidence in a semantics annotation task. WordNet: An Electronic Lexical Database, ed. by Christiane Fellbaum. Cambridge, MA: MIT Press, 217-238.

Gibbs, Raymond W. and Jessica F. Moise. 1997. Pragmatics in understanding what is said. Cognition 62:51-74.

Grice, H. Paul. 1967. Logic and conversation. In Grice 1989:22-40.

Grice, H. Paul. 1989. Studies in the Way of Words. Cambridge, MA: Harvard University Press.

Guasti, Maria Teresa, Gennaro Chierchia, Stephen Crain, Francesca Foppolo, Andrea Gualmini, and Luisa Meroni. 2005. Why children and adults sometimes (but not always) compute implicatures. Language and Cognitive Processes 20(5):667-696.

Hirschberg, Julia. 1991. A Theory of Scalar Implicature. New York, NY: Garland.

Horn, Laurence R. 1984. Towards a new taxonomy for pragmatic inference: Q- and R-based implicature. Meaning, Form, and Use in Context, ed. by Deborah Schiffrin. Washington D.C.: Georgetown University Press, 11-42.

Horn, Laurence R. 1993. Economy and redundancy in a dualistic model of natural language. SKY 1993: 1993 Yearbook of the Linguistic Association of Finland, 33-72.

Huang, Yi Ting and Jesse Snedeker. 2009. Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. Cognitive Psychology 58(3): 376-415.

Huang, Yi Ting and Jesse Snedeker. 2011. 'Logic and Conversation' revisited: Evidence for the division between semantic and pragmatic content in real time language comprehension. Language and Cognitive Processes 26(8): 1161-1172.

Huang, Yi Ting, Elizabeth Spelke, and Jesse Snedeker. 2004. What exactly do numbers mean?, ed. by Ken Forbus, Dedre Gentner, & Terry Regier, Proceedings of the 26th Annual Conference of the Cognitive Science Society, Mahwah, NJ: Erlbaum, 1570.

Huang, Yi Ting, Elizabeth Spelke, and Jesse Snedeker. 2010. When is four far more than three? Children's generalization of newly acquired number words. Psychological Science 21(4): 600-606.

Katsos, Napoleon. 2009. Neither default nor particularised: Scalar implicature from a developmental perspective. Experimental Semantics and Pragmatics, ed. by Uli Sauerland and Kazuko Yatsushiro. Houndsmills, New York, NY: Palgrave MacMillan, 51-73.

Katsos, Napoleon and Dorothy V.M. Bishop. 2011. Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. Cognition. In press. [Available on-line: doi:10.1016/j.cognition.2011.02.015]

Levinson, Stephen C. 2000. Presumptive Meanings: The Theory of the Generalized Conversational Implicature. Cambridge MA: MIT Press.

Nicolle, Steve and Billy Clark. 1999. Experimental pragmatics and what is said: A reply to Gibbs and Moise. Cognition 69:357-354.

Noveck, Ira A. 2001. When children are more logical than adults: investigations of scalar implicature. Cognition 78:165–188.

Noveck, Ira A., Coralie Chevallier, Florelle Chevaux, Julien Musolino, and Lewis Bott. 2009. Children's enrichments of conjunctive sentences in context. Utterance Interpretation and Cognitive Models (Current Research in the Semantic/Pragmatics Interface, Vol. 20), ed. by Philippe De Brabanter and Mikhail Kissine. Emerald Publishing Group, 211–234.

Noveck, Ira A. and Florelle Chevaux. 2002. The pragmatic development of and. BUCLD 26: Proceedings of the 26th Annual Boston University Conference on Language Development, ed. by Barbora Skarabela, Sarah Fish, and Anna H.-J. Do. Sommerville, MA: Cascadilla Press, 453–463.

Noveck, Ira A. and Andres Posada. 2003. Characterizing the time course of an implicature: An evoked potentials study. Brain and Language 85:203–210.

Panizza, Daniele, Yi Ting Huang, Gennaro Chierchia, and Jesse Snedeker. 2011. Relevance of polarity for the online interpretation of scalar terms. Proceedings of Semantics and Linguistic Theory (SALT) 19:360–378.

Papafragou, Anna and Julian Musolino. 2003. Scalar implicatures: Experiments at the semantics-pragmatics interface. Cognition 86(3):253-282.

Papafragou, Anna and Niki Tantalou. 2004. Children's computation of implicatures. Language Acquisition 12(1):71-82.

Pouscoulous, Nausicaa, Ira A. Noveck, Guy Politzer, and Anne Bastide 2007. A development investigation of processing costs in implicature production. Language Acquisition 14(4):347–375.

Récanati, François. 1993. Direct Reference: From Language to Thought. Oxford: Blackwell.

Sperber, Dan and Deirdre Wilson. 1986. Relevance: Communication and Cognition. Cambridge, MA: Harvard University Press.

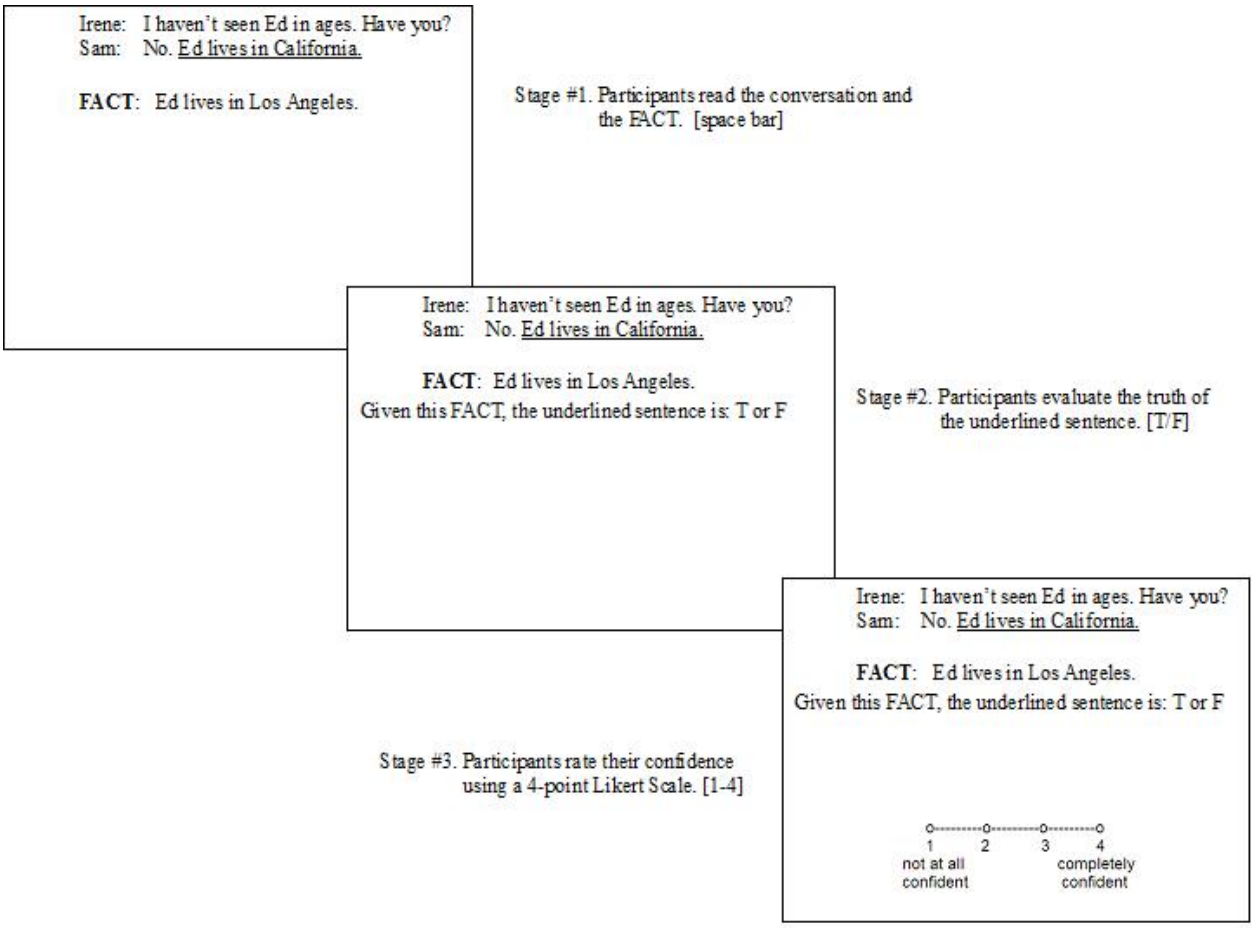

Figure 1. Example of Task Progression

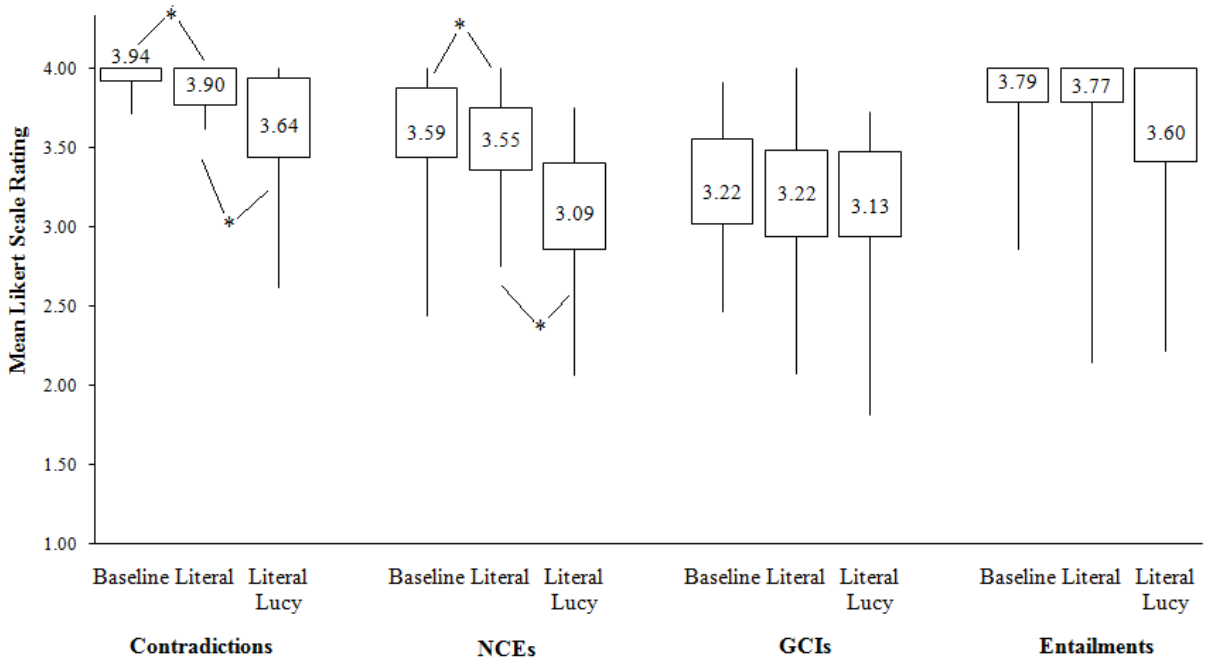

Figure 2. Boxplots of Mean Likert Ratings by Stimulus Type and Instruction Condition

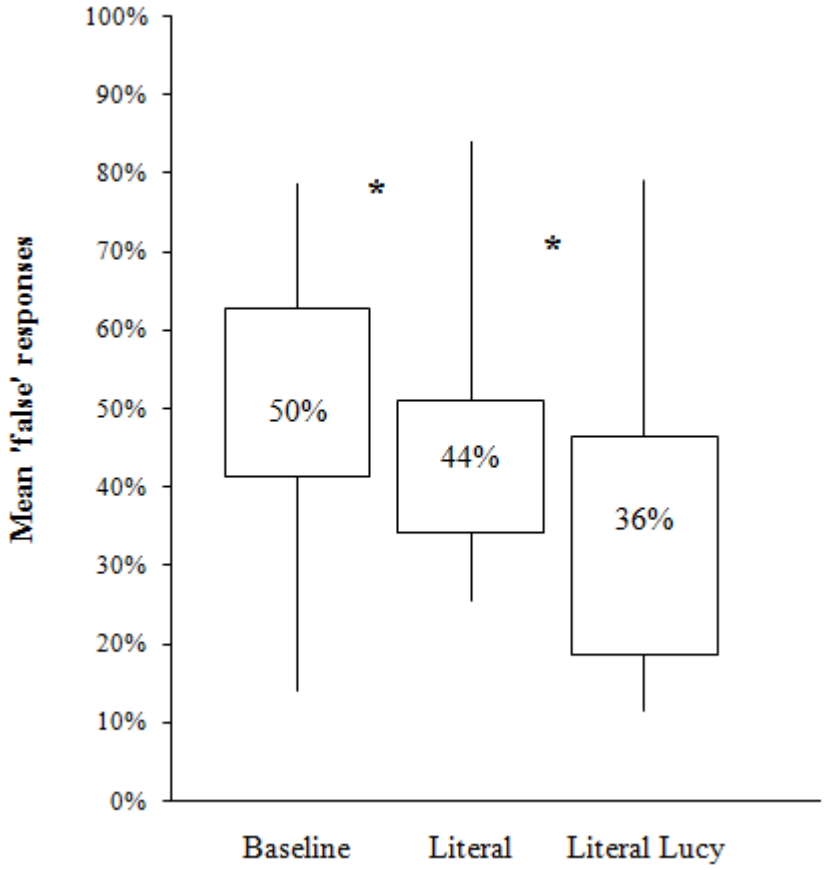

Figure 3. Mean ‘false’ Response Rate by Instruction Condition

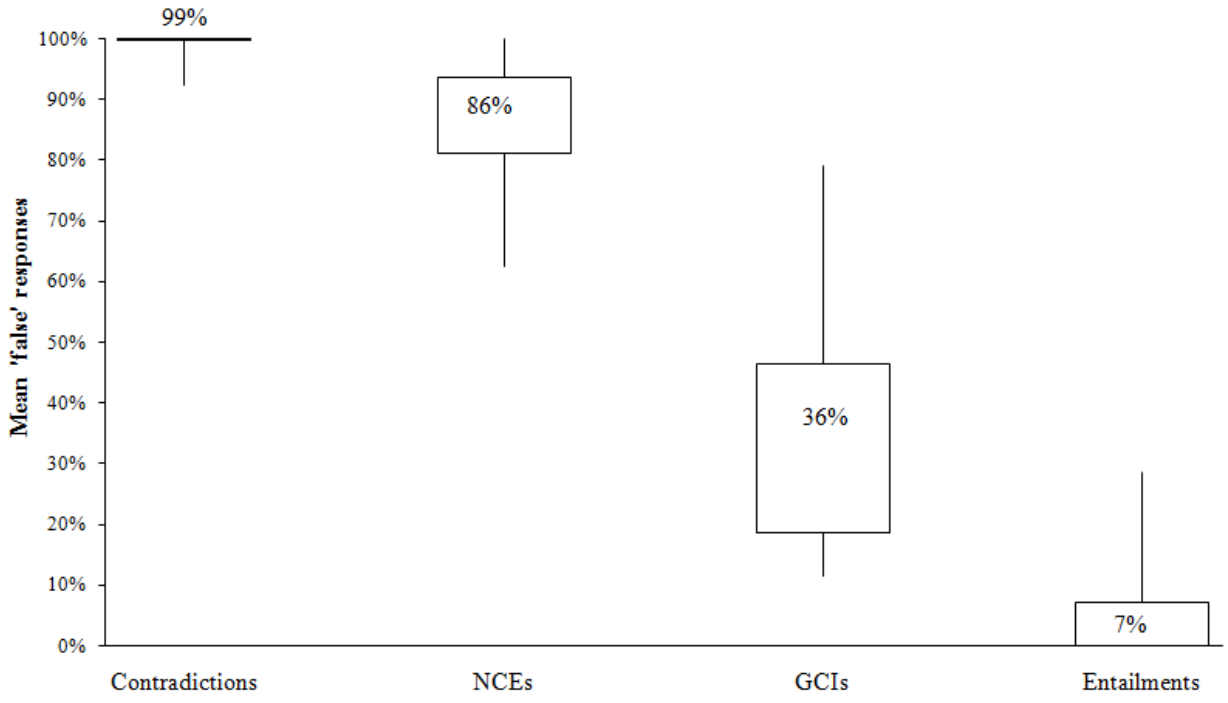

Figure 4. Mean 'false' Responses by Stimulus Type (Literal Lucy Condition only)

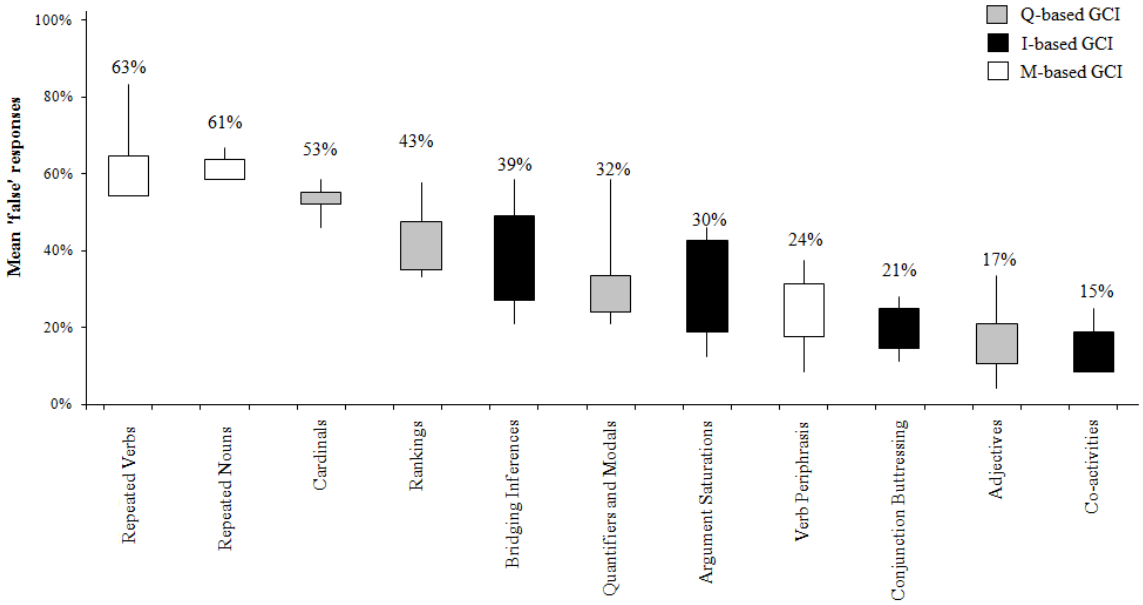

Figure 5. Boxplot of 'false' Responses by GCI Type

[1] We will not, however, attempt to address here questions about whether there may be additional types of NCEs beyond the generally accepted ones of disambiguation, pronoun resolution, and the interpretation of indexicals. For a discussion of other possible NCEs, see Bach (1994).

[2] Post-Griceans have used different terms to refer to pragmatically-determined aspects of utterance meaning that intrude into the truth conditions of an utterance. Sperber & Wilson (1985) and Carston (2002) use explicature, while Récanati (1993) uses pragmatic enrichment. For Post-Griceans, Grice's category of GCIs are not classified as implicatures; for them, such 'implicatures' can constitute part of (truth-conditional) propositional content. For reasons that will become clear below, we continue to use the traditional term 'GCI'.

[3] As the results of our empirical study suggests, cancelling a pragmatically-determined aspect of utterance meaning will depend on a variety of factors, including the particular type of (what Grice called) generalized conversational implicature involved; see below.

[4] Implicatures arising from conjunctions have also been investigated in Noveck & Chevaux 2002 and in Noveck et al. 2009.

[5] One notable exception to this generalization is the case of numeric cardinals. Huang, Spelke, & Snedeker (2004), for example, found that children consistently interpret cardinals as upper-bounded (i.e. with an 'exactly' interpretation). Even in situations with no upper-bounded interpretation available, when children are given a choice between the lower-bounded interpretation of the cardinal (i.e. with an 'at least' interpretation) and an unknown quantity, they will choose the unknown quantity rather than one corresponding to the lower-bounded interpretation.

[6] There is an ongoing discussion in the literature about whether scalar implicature constitutes a unitary phenomenon. Levinson (2000) and Chierchia (2004), for example, suggest that

implicatures that arise from context-independent scales (e.g. cardinals and quantifiers) are privileged compared to those implicatures that arise from context-dependent orderings (e.g. rankings and other ad hoc scales). On the other hand, Horn (1984), Sperber & Wilson (1986), Hirschberg (1991), Carston (2002), Breheny et al. (2006), and Katsos & Bishop (2011), inter alia, argue for a unitary analysis of scalar implicature that does not privilege one type over the other. Regardless of whether scalar implicature constitutes a unitary phenomenon, Doran et al. (2009) show that various factors affect the ease with which a variety of scalar implicatures are incorporated into truth-conditional meaning.

[7] As discussed below, our study included three different instruction conditions. Here, we illustrate the task using the Baseline condition. See §4.2 for details.

[8] A full list of the items used in the experiment can be found at [insert URL].

[9] The taxonomies of Horn (1984) and Levinson (2000) are classificatory in nature, based on the inference mechanisms each takes to be underlying the generation of implicature. As such, their taxonomies are not intended to account for whether – and, if so, the frequency with which – GCIs are incorporated into truth-conditional meaning.

[10] The stimuli were divided into 4 blocks. Each block contained 1 example of each of the 11 types of GCIs used in the experiment, 1 example of each of the 4 types of NCEs, and either 3 Entailments and 4 Contradictions or 4 Entailments and 3 Contradictions. Thus, there was a combined total of 22 stimuli in each block.

[11] A comparison of a main effects model and an interaction model found that the interaction model fit the data better ($\chi^2(6, N = 72) = 66.65$, $p < 0.001$, difference in log likelihood = 33.4).

[12] In a separate logistic regression with dummy coding for both the instruction condition and the stimulus types, we compared the Baseline instruction condition to the Literal Lucy condition and found no significant difference between them (estimate = -0.09, p = 0.31, CI [-0.27, 0.09]).

[13] For the various types of GCIs used in our study, there was no discernible pattern to suggest that confidence ratings differed among them. All GCI types fell within the expected range of confidence ratings; however, we lack sufficient power to test this statistically.

[14] A 'true' response by a participant does not distinguish between whether a given implicature was cancelled or not generated in the first place. Such a distinction, however, is not relevant for our purpose because in both cases participants are not incorporating the GCI associated with the underlined sentence into the truth conditions.

[15] Recall that we are reporting the percentages for only the Literal Lucy condition because this condition best facilitated participants' ability to exclude GCIs from truth conditional meaning. However, for the other instruction conditions, the relative ordering of frequency of incorporation for the different GCI types was preserved: In neither of the other two conditions was the rank order of a GCI type more than two positions away from its rank ordering in the Literal Lucy condition illustrated in Figure 5. Interestingly, for only one GCI type (cardinals) was the difference in frequency of incorporation across instruction conditions statistically significant, with the Baseline condition (92% incorporation) being different from the Literal condition (79%) ($z = 4.45$, $p < 0.001$), and the Literal condition being different from the Literal Lucy condition (53%) ($z = 3.29$, $p < 0.001$).

[16] In response to a concern expressed by an anonymous referee, we performed Chi-square tests to see whether the four items in each of the eleven GCI types patterned similarly. We found that for only two GCI types was there significant deviation from the means: Quantifiers ($\chi^2(3, N =$

24) = 10.05, p < 0.05) and Bridging Inferences (χ^2(3, N = 24) = 8.57, p < 0.05). Both of these GCI types had one experimental item that differed significantly from the others of the same type. For example, the item containing the quantifier most was more likely to be judged false than other items of the same type. However, as only four items were included within each GCI type, we are not able to confidently state that these effects are attributable to the specific linguistic items used.

[17] As with the GCIs, we observed different response rates for the four different NCE types. In the Literal Lucy condition, the percentage of 'false' responses was 95% for indexicals, 93% for ellipses, 89% for deictics, and only 70% for pronouns. As the NCE category was included in our study only to provide a contrast with the GCI types, and the NCE stimuli were not intended to be representative of the broad class of context-dependent elements required for truth-conditional meaning, we must be cautious about drawing unwarranted conclusions based on these findings. (See §6 below.)

[18] More recently, studies of the online processing of scalar implicatures (evoked by utterances including quantifiers, such as some, but not cardinals) suggest that truth-conditional content is generated prior to implicated content, supporting the notion that the two are separately represented, although the latter follows rather rapidly in online processing (Panizza et al. 2011).

[19] One possibility, of course, is the Post-Gricean position that there is simply no principled distinction between particularized and generalized conversational implicature. That is, one might argue that there are no theoretically relevant generalizations to be made about the class of GCIs that hold of GCIs but not also of PCIs. Under this view, it would follow that so-called GCIs do not warrant a theoretically privileged role in theories of meaning, an interesting position that we will not be pursuing here.