PSYCHOLINGUISTIC AND NEUROLINGUISTIC INVESTIGATIONS OF SCALAR IMPLICATURE

By

Stephen Politzer-Ahles

Submitted to the graduate degree program in Linguistics and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Chairperson Robert Fiorentino

Alison Gabriele

Susan Kemper

Utako Minai

Jie Zhang

Date Defended: 23 April 2013

The Dissertation Committee for Stephen Politzer-Ahles

certifies that this is the approved version of the following dissertation:

PSYCHOLINGUISTIC AND NEUROLINGUISTIC INVESTIGATIONS OF SCALAR IMPLICATURE

Chairperson Robert Fiorentino

Date approved: 23 April 2013

ABSTRACT

The present study examines the representation and composition of meaning in scalar implicatures. Scalar implicature is the phenomenon whereby the use of a less informative term (e.g., *some*) is inferred to mean the negation of a more informative term (e.g., to mean *not all*). The experiments reported here investigate how the processing of the implicature-based aspect of meaning (e.g., the interpretation of *some* as meaning *not all*) differs from other types of meaning processing, and how that aspect of meaning is initially realized.

The first three experiments measure event-related potentials (ERPs) to examine whether inferential pragmatic aspects of meaning are processed using different mechanisms than lexical or combinatorial semantic aspects of meaning, and whether inferential aspects of meaning can be realized rapidly. Participants read infelicitous quantifiers for which the semantic meaning (*at least one of*) was correct with respect to the context but the pragmatic meaning (*not all of*) was not, compared to quantifiers for which the semantic meaning was inconsistent with the context and no additional pragmatic meaning is available. Across experiments, quantifiers that were pragmatically inconsistent but not semantically inconsistent with the context elicited a broadly distributed, sustained negative component. This sustained negativity contrasts with the N400 effect typically elicited by nouns that are incongruent with their context, suggesting that the recognition of scalar implicature errors elicits a qualitatively different ERP signature than the recognition of lexico-semantic errors. The effect was also distinct from the ERP response elicited by quantifiers that were semantically inconsistent with a context. The sustained negativity may reflect cancellation of the pragmatic inference and retrieval of the semantic meaning. This process was also found to be independent from lexico-semantic processing: the N400 elicited by lexico-semantic violations was not modulated by the presence of a pragmatic inconsistency. These findings suggest there is a dissociation between the mechanisms for processing combinatorial semantic meaning and those for inference-based pragmatic meaning, that inferential pragmatic meaning can be realized rapidly, and that the computation of meaning involves continuous negotiation between different aspects of meaning.

The next set of experiments examined how scalar implicature-based meanings are realized initially. Default processing accounts assume that the interpretation of *some of* as meaning *not all of* is realized easily and automatically (regardless of context), whereas context-driven processing accounts assume that it is realized effortfully and only in certain contexts. In two experiments, participants' self-paced reading times were recorded as they read vignettes in which the context did or did not bias the participants to make a scalar inference (to interpret *some of* as meaning *not all of*). The reading times in the first experiment suggested that the realization of the inference was influenced by the context: reading times to a target word later in the vignette were facilitated in contexts in which the scalar inference should be realized but not in contexts where it should not be realized. Importantly, however, reading times did not provide evidence for processing cost at the time the inference is realized, contrary to the predictions of context-driven processing accounts. The results raise the question of why inferencing occurs only in certain contexts if it does not involve extra processing effort. In the subsequent experiment, reading times suggested that the inference may not have been realized when participants engaged in a secondary task that increased processing load. These results, together with the results of other recent experiments, suggest that inferencing may be effortless in certain contexts but effortful with other contexts, and not computed at all in still other contexts, depending on the strength of the bias created by the context. These findings may all be accountable for under a recently-proposed constraint-based processing model of scalar implicature.

ACKNOWLEDGEMENTS

I would like to thank my dissertation committee for their feedback on this research, and in particular to thank my chair, Robert Fiorentino, for his guidance and collaboration throughout all of the research described in this dissertation, and for giving me the opportunity and the tools to conduct this research.

The research described in Chapter 2 was conducted in collaboration with Xiaoming Jiang and Xiaolin Zhou at the Center for Brain and Cognitive Sciences at Peking University, and I am grateful for their cooperation and feedback. Some of the projects were funded by the NSF's East Asia and Pacific Summer Institutes program, without which I would not have had the opportunity to collaborate with this laboratory. I received additional support from the University of Kansas Doctoral Student Research Fund to travel to China for data collection. Many of the materials for these experiments were created or edited for me by Yan Liang, Chunping Wu, Yue Wu, Yingyi Luo, Mengyan Zhu, Junru Wu, Lu Yang, and Lamar Hunt III; Yin Wu and Yuqin Zhou provided much assistance with data collection.

I thank Kelly Berkson and Natalie Pak for their assistance in constructing the materials for the research described in Chapter 3, and Patrick Patterson for his assistance with data collection. This research was partly funded by an NIH T32 institutional grant.

This research has benefitted from the feedback of numerous colleagues at the Cognitive Neuroscience Society conference (2011 and 2012), Neurobiology of Language conference (2011 and 2012), and CUNY Conference on Human Sentence Processing (2013). It has also benefitted greatly from the feedback of several journal editors and anonymous reviewers. I am also grateful for the frequent feedback from my colleagues in the University of Kansas Linguistics Department, particularly the participants in the Research in Acquisition and Processing seminar,

the Research in Experimental Linguistics seminar, the Child Language Proseminar, and the Linguistics Colloquy.

Thankfully scalar implicature was only how I spent some, and not all, of my time here in Kansas, so I would also like to express my gratitude to my friends and Linguistics cohorts here, in no particular order: Kelly Berkson, Joshua Shireman, Bea Lopez, Jose Aleman-Banon, Mircea Sauciuc, Turki Binturki, Breanna Steidley, Lamar Hunt III, Jiang Liu, Yuanliang Meng, Hyunjung Lee, Goun Lee, Phil Duncan, Maite Martinez-Garcia, Le Ann Swallom, Joleen Chu, Kate Coughlin, Travis Major, Jon Coffee, Hiba Gharib, Mahire Yakup, Grace Zhang, Ibrahima Ba, Ethan Skinner, Robert Lewis, Midam Kim, Yuka Naito-Billen and Sam Billen, Terry Hsieh, Juwon Lee, Tom Dirth, the KU Badminton Club, the Sandrats, and whoever else I may have forgotten to mention.

Finally I would like to thank my parents—Geneva, Ben, and Marilyn—for their support, for their understanding of my choice to pursue the weird life of an academic, and for their patience in the face of all the uncertainties that come along with that.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Comprehending language involves composing meaning out of multiple units: the meaning of a sentence like "The cat sat on the mat" is composed out of the meanings of the individual words in the sentence and the relationships between these words. Units of meaning may also come from outside the sentence. For instance, based strictly on its words and grammar the sentence "Can you pass the salt?" is a question about someone's abilities, but someone hearing that sentence at the dinner table would interpret it instead as a request to give the salt to the speaker. This interpretation is based on an inference about the speaker's intentions (Grice, 1975): the person uttering that sentence probably is not questioning the hearer's abilities but probably does care about having the salt, and thus the hearer infers what the speaker meant, even though that meaning is not included in the literal semantics of the sentence that was uttered. In short, comprehension of even simple utterances involves integration of different aspects of meaning coming from both within the utterance and from expectations about other people's intentions.

Language users perform this sort of integration ubiquitously and without apparent effort (Van Berkum, 2009). Thus, to understand language comprehension, we must also understand how multiple aspects of meaning are realized, compared, and combined or rejected during processing. People are rarely consciously aware of how they are composing meaning as language is unfolding; therefore, these processes can usually only be measured via implicit online measures that are sensitive to cognitive processes taking place prior to and independently of overt responses or decisions. Furthermore, meaning is composed rapidly (there is little delay between hearing a sentence like "Can you pass the salt?" and understanding what the appropriate response is, and only in exceptional circumstances do people engaged in conversation need to stop and think before understanding the meaning of a simple sentence) and incrementally (comprehenders do not wait until the end of a sentence to start putting together the meanings of words and inferring speaker meaning). Therefore, the cognitive processes underlying meaning composition must be investigated using methods with high temporal resolution that can reveal these processes as a sentence is unfolding—and precisely where and when in a sentence these processes occur—rather than methods that only consider how the sentence is ultimately interpreted. The studies presented in this dissertation use implicit, online measures of this type to investigate the cognitive mechanisms underlying inference and meaning composition, focusing on a particular linguistic phenomenon, *scalar implicature*.

1.1. SCALAR IMPLICATURE: LINGUISTIC THEORY

Scalar implicature refers to the interpretation of a weak (less informative) term as meaning that a stronger (more informative) term does not hold.[1] Consider, for instance, the exchange in (1):

1) A. Are all of the students in your department hardworking?
   B. Some of them are.

In this context, because speaker B chose not to say "All of them are", a hearer often interprets the utterance *some of them are* as meaning *not all of them are*. This interpretation is considered "scalar" because the quantifiers *some of* and *all of* are assumed to occupy a lexical scale, <*some*, *all*>, in which both express the same sort of information (the quantity of elements in some set have some property, such as being hardworking), but *all of* is the "stronger" element of the scale in that it makes a stronger, more informative claim—there are fewer possible scenarios in which

[1] Throughout this dissertation, both the terms *scalar implicature* and *scalar inference* will be used to refer to the phenomenon being discussed. The term *scalar implicature* will be used to refer to the act on the speaker's/utterer's part (e.g., when discussing how a comprehender "comprehends a scalar implicature"), and *scalar inference* will be used to refer to the act on the comprehender's part (e.g., when discussing how a comprehender "makes a scalar inference" or "realizes a scalar inference").

all of the X can be true (Noveck & Sperber, 2007; Levinson, 2000; Grice, 1975; Horn, 1972). Other kinds of linguistic expressions are also thought to occupy lexical scales—e.g., coordinators (*X or Y* may be interpreted as meaning *not [X and Y]* because of the lexical scale <*or*, *and*>), adjectives (*warm* may be interpreted as *not scalding* because of a lexical scale such as <*warm*, *hot*, *scalding*>, etc.), and more—although different expressions and different scales differ in the strength of the scalar implicature they invoke (Doran, Baker, McNabb, Larson, & Ward, 2012; Doran, Ward, Larson, McNabb, & Baker, 2009).

According to Gricean accounts of scalar implicature, the interpretation due to scalar inference—i.e., the interpretation of *some of* as meaning *not all of*—is a pragmatic meaning based on an inferential enrichment process, and is not part of the inherent semantics of the quantifier *some of* (Noveck & Sperber, 2007; Grice, 1975; Horn, 1972). The pragmatic interpretation arises based on a hearer's expectation that a cooperative speaker will use the most informative expression possible—i.e., if *all of* were true, then the speaker would have said *all of* rather than the less informative expression *some of*. Thus, when the speaker chooses not to use the stronger expression, the hearer infers that the stronger expression must not be true—i.e., that *some of* must mean *not all of*.

Crucial to the notion of scalar implicature, a term like *some of* has a semantic interpretation that is separate from the inference-based, pragmatic interpretation. The semantic interpretation of *some of*, for example, is the existential (*at least one of*) used in logic and syllogistic reasoning (Newstead, 1988). In other words, under the semantic interpretation of the expression, the possible scenarios in which *all of the X are Y* is true are a subset of the possible scenarios in which *some of the X are Y* is true (because if all of the students are hardworking, then any random group of "some of them" must also be hardworking). Under the pragmatic interpretation, on the other hand, the set of possible scenarios in which *all of the X are Y* is true is disjoint from the set of possible scenarios in which *some of the X are Y* is true (since *some of* means *not all of* under this interpretation, *some of* and *all of* cannot be true in the same scenario). A common argument for the distinction between the pragmatic and semantic interpretation is the *defeasibility* (or *cancellability*) argument: the pragmatic interpretation can be cancelled or revised (as in (2a) below) without resulting in a nonsensical sentence (Doran et al., 2012; Rullman & You, 2006), whereas the semantic meaning cannot (as in (2b)):

2) a. Some of the students in this department are hardworking. In fact, all of them are.
   b. Some of the students in this department are hardworking. #In fact, none of them are.

Thus, the *not all of* interpretation of *some of* is thought to be an inference, and the *at least one of* interpretation to be an entailment. In (2a), the inference *not all of the students in this department are hardworking* is explicitly cancelled by the following sentence, which specifies that all of the students are; while this cancels the inference, it does not result in a contradiction, indicating that the comprehender must be able to re-interpret *some of* semantically as meaning *at least one (and possibly all)*. In (2b), the second sentence instead cancels the entailment *at least one of the students in this department is hardworking*, resulting in a contradiction. (See, however, Meibauer, 2012, for a review of challenges to the defeasibility argument.) The pragmatic interpretation is also known as the *upper-bounded* interpretation (because it asserts that the upper bound *all of*, the largest possible set of *X*s that are *Y*, is not true), and the semantic interpretation as the *lower-bounded* interpretation (because it asserts that the lower bound *none of*, the smallest possible set of *X*s that are *Y*, is not true; *at least one of* could be rephrased as *not none of*).

A related piece of evidence for the distinction between pragmatic and semantic interpretations comes from cases in which the inference is not cancelled by a later utterance, but seems to not arise at all in a particular context, as in (3) (adapted from Levinson, 2000):

3) A. Was there any evidence against them?
   B. Yes, some of their documents were forgeries.

It is generally assumed (Katsos & Cummins, 2010: 285; Levinson, 2000: 51) that B's utterance in this example is not interpreted as meaning *it is not the case that all of their documents were forgeries*, since that information is not relevant to A's question. Again, this is taken as evidence that the inference-based *not all of* interpretation is separate from the semantics of *some of*. Other contextual and linguistic factors that contribute to whether *some of* is interpreted semantically include the presence or absence of lexical alternatives in the context (the inference may be inhibited when numbers like *two* and *three* were possible alternatives to *some of* in the context; Degen & Tanenhaus, 2011; Huang, Hahn, & Snedeker, 2010), the syntactic form of the expression itself (i.e., partitive *some of* is interpreted pragmatically more often than *some*; Degen & Tanenhaus, 2011), prosody (contrastive stress on *some* makes the inference more likely, as does reduction of *some of* into *summa*; Degen & Tanenhaus, submitted; Grodner, Klein, Carbary, & Tanenhaus, 2010), and syntactic position/information structure (in Greek, the quantifier corresponding to *some of* is more likely to be interpreted pragmatically when it is in sentence-initial subject position, which is associated with given/old information; Breheny, Katsos, & Williams, 2006). Whether these examples of previous context inhibiting a pragmatic interpretation are actually different from the earlier example of later context cancelling a pragmatic interpretation remains an open question; some psycholinguistic accounts of inferencing argue that in all of these examples the pragmatic inference is realized automatically at first, but cancelled before the comprehender is aware of it. This question will be discussed in the following section, and forms the basis for the experiments discussed in Chapter 3.

Since scalar implicatures introduce a dissociation between semantic and pragmatic meaning, since their realization can be manipulated with minimal changes in context (as will be shown in Chapter 3), and since a comprehender's interpretation of an expression like *some of* can be tested using implicit online methods, scalar implicatures offer an ideal test case for examining the relationship between semantics and pragmatics in the dynamics of sentence comprehension. It is not universally accepted, however, that the "pragmatic" interpretation of scalar expressions like *some of* is actually pragmatic. The grammatical view[2] of scalar implicatures (Chierchia, Fox, & Spector, 2012; Chierchia, 2004; for more reviews and discussion, see Geurts & van Tiel, to appear; Chemla & Spector, 2011; Ippolito, 2011; and Geurts & Pouscoulous, 2009; inter alia) holds that they do not arise from a pragmatic inference, but rather that they arise as part of semantic composition via the insertion of an "exhaustification" operator (similar to a covert version of *only*, changing *some of* into *only some of*). The operator is inserted in contexts where its insertion would lead to a stronger (more informative) expression. This view has been argued based on the fact that the realization of scalar implicature interacts with the scope of other semantic operators such as polarity items. Under such a view, scalar implicature is a semantic rather than a pragmatic, inference-based phenomenon. Nevertheless, it is still the case that the upper-bounded *not all of* interpretation of *some of* seems to have a different status than the lower-bounded *at least one of* interpretation. Whether this difference is a difference between pragmatic and semantic meaning, or a difference between different types of semantic meaning, is an important question, but will not be addressed in this dissertation. The questions that will be raised in the following sections—questions of whether the two types of meaning are processed differently, and how the upper-bounded meaning is realized—are questions that are pertinent to developing and adjudicating between psycholinguistic models of meaning realization regardless of whether we take the upper-bounded meaning to be semantic or pragmatic in nature.

[2] This is also sometimes referred to as a *localist* view, as opposed to *globalist* theories which assume the inference is a pragmatic phenomenon (like the view first described in this section).

It should be noted that, even when *some of* is given the upper-bound or lower-bound interpretation, it is not the case that all values are equally acceptable—that is to say, in a situation where a comprehender interprets *some of* as meaning *not all*, the comprehender may still believe the quantifier is less acceptable for describing a situation in which four out of six elements in a set (for example) meet some condition than a situation in which two out of six elements meet that condition (see Degen & Tanenhaus, 2011; inter alia). The interpretation of a quantifier like *some of* can be characterized in terms of *fuzzy set theory*, such that possible values (i.e., sizes of subsets that may be referred to by the quantifier) are not always wholly within or outside the meaning of the quantifier, but rather may be partially within the meaning of the quantifier (Newstead, 1988). In fuzzy set theory, the extent to which a given value fits within the range of the quantifier is indicated by the quantifier's *membership function*, the output of which is a value between 0 and 1. For example, in a set of 13 items (such as those tested by Degen & Tanenhaus, 2011), a subset consisting of 5 items may be definitely within the range that could be described by *some of* (and receive a value of 1 from the membership function), whereas an empty subset consisting of 0 items may be definitely outside the range (and receive a value of 0), whereas a subset consisting of all 13 items may receive an intermediate value from the membership function. Furthermore, the range of values which may be felicitously described using *some of* can be shifted by numerous aspects the context, including set size, expectations, and the presence of alternative quantifiers in the context; for reviews, see Newstead (1988), Noveck and Sperber (2007), and Degen and Tanenhaus (2011). This dissertation will focus specifically on the processes involved in introducing the *not all* upper bound in the interpretation. It remains an empirical question whether those processes are the same as processes that modify other aspects of the range of acceptable values (the membership function) for *some of*.

1.2. SCALAR IMPLICATURE: PSYCHOLINGUISTIC MODELS

The literature and arguments reviewed above strongly suggest that scalar expressions like *some of* have two different interpretations and that these interpretations enjoy a different status in terms of their defeasibility. A major question remaining is how are these different meanings processed, and how is the upper-bounded, "pragmatic" meaning realized by comprehenders during online language process? Several psycholinguistic accounts (none of which have been computationally implemented, to my knowledge) have been proposed in response to this question. These can be broadly described as *context-driven*, *default*, and *constraint-driven* accounts, although each of these classes of accounts can be formalized into various different models (see, e.g., Bott, Bailey, & Grodner, 2012, for several possible types of default models).

Context-driven models. As described in the previous section, the realization of a scalar inference (i.e., the interpretation of *some of* as meaning *not all of*) is thought to involve an extra process beyond that of realizing the expression's semantic meaning—the comprehender must realize that the speaker had other, stronger expressions available to her, and must make the inference that if the speaker did not use one of those expressions then she must have meant to indicate that they were not true. Context-driven models of scalar inferencing assume that these operations require processing effort. The context-driven accounts are based on Relevance Theory (Sperber & Wilson, 1995) and adopt one of its central tenets: that the parser does not undergo cognitively costly operations unless it has something to gain (in terms of the specificity of the information communicated) by doing so. Thus, the assumption that inferencing requires effort leads to two related predictions. First, context-driven models predict that a scalar inference will only be realized in contexts where the upper-bounded interpretation (e.g., *not all of*) is relevant to the discourse—hence the name "context-driven". Secondly, they predict that the upper-bounded interpretation will be realized after the lower-bounded, semantic interpretation—it will be delayed by at least as much time as it takes the parser to determine whether the context is one in which the inference is worthwhile, plus the amount of time it takes to actually realize the inference.

Context-driven models, like default models, are "theories of linguistic representation and not of language processing" (Huang & Snedeker, 2009: 408), and thus their predictions are not explicit about all aspects of processing, such as how long it takes the parser to conduct the Gricean reasoning described in the previous section or precisely what processing components are taxed by inferencing. Nevertheless, their predictions are still quite different than those of default models, described below.

Default models. These models, also referred to as *neo-Gricean* models, instead assume that realizing certain kinds of inferences[3] is rapid and cost-free. Such models, which are due mainly to Levinson (2000; see also Gadzar, 1979, and Horn, 1984), are based on the idea that the enriched interpretation of a term like *some of* is the preferred and more commonly used meaning in natural language. Because they are so often used, the language parser is argued to have developed heuristics to facilitate rapid communication. Any time a scalar expression like *some of* is uttered, the upper-bounded meaning is evoked without regard to the linguistic and discourse context; the linguistic form of the quantifier itself is sufficient to evoke the inference. If the inference is then shown to be contextually inappropriate or unnecessary, it may be cancelled (Levinson, 2000). Thus, whereas context-driven accounts predict that the inference will become available *only if* the context supports it, default accounts predict that the inference will become

[3] Specifically, Generalized Conversational Implicatures (GCIs), as opposed to Particularized Conversational Implicatures (PCIs). In these accounts, the inference evoked by *some of* is a GCI. As the present dissertation only discusses GCIs like *some of*, Grice's proposed distinction between GCIs and PCIs is beyond the scope of the present work, and in fact Relevance theory does not accept that there even is such a distinction. For further discussion of the GCI/PCI distinction see Levinson (2000), Katsos & Cummins (2010) and Breheny and colleagues (in press), among others.

available *unless* the context does not support it. Even in cases where the context does not support the inference, default accounts predict that the inference will be realized briefly and then cancelled, whereas context-driven accounts predict that the inference will not be realized at all. Levinson does not make specific predictions about the amount of time that the inference cancellation process takes (and by extension, the amount the inference would remain temporarily available during processing) or the processing costs of such a process. In their review, Katsos and Cummins (2010) assume that this cancellation procedure should require processing time and resources.

Constraint-based model. A recent proposal by Degen and Tanenhaus (2011, submitted) follows constraint-based accounts of syntactic and semantic comprehension (e.g., Trueswell, Tanenhaus, & Garnsey, 1994) in assuming that the parser evaluates all available information as early as possible and uses this information to facilitate or inhibit the scalar inference. Thus, unlike the previous accounts which assume multiple stages (either a first stage in which semantic meaning is used before context has been evaluated, as in context-driven models, or a first stage in which generalized conversational implicatures are realized automatically, as in default models), this model proposes that all information influences scalar inference realization immediately. Thus, in situations where numerous constraints that have already been processed (such as discourse and semantic context) to facilitate a scalar inference, the inference may be realized rapidly and effortlessly; in situations where few constraints facilitate the inference, or where constraints actively discourage it, the inference may be realized slowly and effortfully or not at all. This constraint based account might be considered a special case of context-driven accounts, given that it assumes scalar inference is dependent on context; it differs from those accounts, however, in that it does not predict scalar inference realization to always be slow and effortful.

The grammatical account of scalar inference proposed by Chierchia and colleagues (Chierchia et al., 2012; Chierchia, 2004) is sometimes treated as a separate account from those described above. It is assumed by Katsos and Cummins (2010) and Huang and Snedeker (2009) to make processing predictions that are similar to those in Levinson's (2000) default account. In particular, the grammatical account assumes that scalar inferences are realized by default (at least in the right entailment contexts) as a result of the linguistic form of the expression. The derivation of the inference, however, is thought to take place through a series of semantic operations, the psychological nature of which are not known. Thus, it is not necessarily clear what the processing predictions of such a model should be (see Panizza, Huang, Chierchia, & Snedeker, 2011, for further discussion).

1.3. THE STRUCTURE OF THIS DISSERTATION

The experiments described in the rest of this dissertation use neurolinguistic and psycholinguistic methods to investigate aspects of scalar implicature processing discussed above. The first series of experiments examines whether event-related brain potentials (ERPs) provide evidence for a difference between inference-based quantifier processing and semantic quantifier processing. Data from three experiments shows that violations of inference-based meaning yield different ERPs than violations of semantic meaning and that these two processes may be functionally independent. The second series of experiments tests the models described above by using a self-paced reading task to examine whether the realization of inference-based meaning entails a processing cost and is sensitive to context. Results suggest that inferencing is indeed context-sensitive but that it does not evoke a directly observable processing cost, which raises challenges for the traditional accounts of scalar inference processing described above.

CHAPTER 2: ELECTROPHYSIOLOGICAL CORRELATES OF INFERENTIAL VERSUS SEMANTIC PROCESSING

2.1. INTRODUCTION[4]

As described in the previous chapter, the comprehension of scalar implicatures involves processing multiple aspects of meaning: a lower-bounded meaning which is semantically inherent to the expression, and an upper-bounded meaning which may be realized through additional pragmatic or semantic processes. A number of recent psycholinguistic studies have investigated the speed at which pragmatic readings of scalar terms become available, the costs engendered by inferencing, and the role of context in scalar implicature processing (see, e.g., Noveck & Posada, 2003; Bott & Noveck, 2004; Feeney, Scafton, Duckworth, & Handley, 2004; Breheny, Katsos, & Williams, 2006; De Neys & Schaeken, 2007; Chevallier, Noveck, Nazir, Bott, Lanzetti, & Sperber, 2008; Degen, 2009; Dieussaert, Verkerk, Gillard, & Schaeken, 2011; Bott et al., 2012; Hartshorne & Snedeker, submitted). Many of these studies have used speeded verification or self-paced reading tasks. Response times in such tasks, however, may reflect not only processing related to implicature generation but also controlled decision-making components (Huang & Snedeker, 2009; Nieuwland et al., 2010; Tavano, 2010). This leaves open the question of what occurs before an overt response (or decision to move to the next word) is made, and how implicature processing unfolds over time. Thus, it is worthwhile to investigate these questions using a methodology that both provides fine-grained temporal resolution and allows the researcher to track different processing stages prior to overt responses.

[4] Portions of this chapter are adapted from Politzer-Ahles, Fiorentino, Jiang, & Zhou (2013) and from Politzer-Ahles, Jiang, Fiorentino, & Zhou (2012).

One such methodology is event-related potentials (ERPs). In addition to offering high temporal resolution, ERPs have the potential to probe the extent to which the neural mechanisms of scalar implicature processing differ from those of other aspects of meaning composition, since ERP components may differ in terms of topography, polarity, and morphology, as well as latency (see, e.g., Kutas et al.,2006). This makes ERPs a particularly useful tool for investigating the interplay between these different aspects of meaning.

2.1.1. Context and pragmatics in ERP studies

Many previous neurolinguistic studies examining pragmatic meaning have focused on real-world plausibility (e.g., Kuperberg et al. 2000; Hagoort, Hald, Bastiaansen, & Petersson, 2004; Filik & Leuthold, 2008), rather than aspects of meaning based on inferential pragmatics—i.e., meaning based on assumptions about the intentions of the speaker who makes an utterance and the context in which she utters it. The experiments reported in this chapter aim to investigate how the brain realizes linguistically-motivated distinctions between different aspects of meaning (semantically inherent meanings versus enriched meanings that are generated through additional pragmatic or semantic processes) and how these aspects of meaning are composed online.

It is well known that information from the wider discourse and pragmatic context is used rapidly during sentence comprehension to make words easier or more difficult to integrate into the utterance meaning (Hagoort & Van Berkum, 2007; Van Berkum, 2009). Pragmatic and discursive information can guide comprehenders' predictions about upcoming words and thus, in ERP studies, produce modulations in the N400, a negative-going ERP component emerging between about 200 and 500 ms after the presentation of a word and showing a greater amplitude to words that are less expected and more difficult to retrieve or integrate (Kutas & Federmeier, 2000; Lau, Phillips, & Poeppel, 2008; Pylkkänen, Brennan, & Bemis, 2011). Previous studies have shown that discourse context can override semantic constraints, making semantically appropriate but discursively inappropriate words elicit an increased N400, an effect normally elicited by semantically anomalous words (Nieuwland & Van Berkum, 2006; Filik & Leuthold, 2008). Language-external variables like the hearer's personal values or the speaker's gender, age, or class can make words easier or more difficult to retrieve from memory and integrate into a sentence and thus influence the N400 (Van Berkum, 2009) and brain activation in the medial prefrontal cortex (Tesink et al., 2009). N400-like ERP responses to pronouns are affected by the social status of their antecedents (Jiang et al., 2011) and gender stereotypes held by the comprehender (Osterhout, Bersick, & McLaughlin, 1997). Pragmatic information can also play a role in semantic composition: there is evidence that negatives are not always rapidly integrated into the meaning of infelicitous sentences such as "A robin is not a bird" (Fischler, Bloom, Childer, Roucos, & Perry, 1983; Wiswede, Koranyi, Müller, Langner, & Rothermund, in press; but see Urbach & Kutas, 2010) but that they are when pragmatic context makes the sentence felicitous (Nieuwland & Kuperberg, 2008).

In contrast to these studies examining how pragmatic context influences retrieval and integration of a later word in the sentence, comparatively few have probed for ERP activity directly related to pragmatic inferencing or tested whether this activity is qualitatively distinct from that elicited by semantic retrieval and integration. Pragmatic inferencing may elicit sustained negativities rather than N400s. A sustained negativity known as the Nref, which begins at a latency of about 300ms in response to words with multiple or ambiguous referents as compared to words with unique referents (Van Berkum, Koornneef, Otten, & Nieuwland, 2007), has been suggested to be related to computationally costly inference-making (Van Berkum, 2009). This hypothesis remains to be tested empirically. Crucially, similar sustained negativities have been observed for sentences in which the reader must re-compute a discourse model about whether or not an action was completed (Baggio, van Lambalgen, & Hagoort, 2008) or revise a discursive inference that turns out to be incorrect (Pijnacker, Geurts, van Lambalgen, Buitelaar, & Hagoort, 2011), although in the latter study the negativity had a more centro-parietal distribution.

2.1.2. ERP studies of scalar inference

To date, only three ERP studies have investigated scalar implicature processing in particular. These studies (Noveck & Posada, 2003; Nieuwland, Ditman, & Kuperberg, 2010; Hunt, Politzer-Ahles, Gibson, Minai, & Fiorentino, 2013) have all examined N400 responses downstream of the scalar expression. These are briefly summarized below.

Noveck and Posada (2003) measured ERPs while participants read and judged *underinformative* sentences such as "Some dogs have ears." Such sentences are correct under a lower-bounded interpretation (there do exist dogs that have ears) but incorrect under an upper-bounded interpretation (it is not the case that "not all dogs have ears"). ERP responses to these sentences were compared to responses to true, informative sentences (e.g., "Some gardens have trees") and false sentences (e.g., "Some toads have churches"). At the sentence-final critical word which determines the truth, falsehood, or underinformativeness of the sentence, the investigators found a decreased N400 for underinformative sentences relative to true sentences. The interpretation of this finding is complicated, however, by between-item differences in lexico-semantic relatedness between subjects and objects in their materials (i.e., the nature of the relationship between "dogs" and "ears", or other subject-object pairs used in the other underinformative sentences, is not the same as the nature of the relationship between "gardens" and "ears", or other subject-object pairs used in other true sentences), the fact that critical words were not matched for any lexical properties (e.g., frequency), and the possible effect of global wrap-up processes that occur at the end of a sentence (for a review of these concerns, see Nieuwland et al., 2010; for a discussion of sentence wrap-up effects, see Hagoort, 2003).

A later study by Nieuwland, Ditman, and Kuperberg (2010, Experiment 1) tested similar sentences, using critical words that were matched for length and frequency and not presented in sentence-final position. Examples of their stimuli are "Some people have pets, which require good care" (true and informative), and "Some people have lungs, which require good care" (underinformative). They also had participants read the sentences passively rather than make judgments, as they were concerned that a judgment task such as that used by Noveck and Posada (2003) could elicit decision-related components which would mask other effects in the N400 time window. In this experiment, the authors found that participants with high pragmatic ability (as measured by performance on the communication subscale of the Autism-Spectrum Quotient questionnaire) showed a greater N400 for underinformative than informative critical words. These results suggest that scalar implicatures can guide expectations about upcoming linguistic input and can override lexico-semantic influences on the N400. Interestingly, in a separate experiment, when the critical words were temporarily underinformative but were followed by restricting relative clauses that made the sentences informative (e.g., "Some gangs have members that are really violent"), the N400 effect was not observed, suggesting that the lack of truly underinformative sentences in the global experimental context modulated participants' brain responses to temporary ambiguity.[5]

[5] Note that this is a different claim than the claim in Section 1.1 that various contextual factors influence whether or not a scalar inference is realized. In this experiment the suggestion is not that the global experimental context made participants not realize the inference, but rather that it made the interpretation of the sentence remain informative even when the inference was realized.

While this experiment controlled many of the potentially problematic factors that were in Noveck and Posada's (2003) experiment, some systematic differences between the underinformative and control sentences remained. Particularly, direct objects in underinformative sentences tended to have a closer semantic relationship to the subjects than did the direct objects in control sentences (compare "Some gangs have members" versus "Some gangs have initiations"). Indeed, the participants with low pragmatic ability tended not to show an increased N400 in response to underinformative critical words, but a decreased one. The authors suggest that whereas the high-ability participants were focusing on the overall meaning of the sentence, the low-ability participants were strategically focusing on the lexical relationships between words.

A later collaborative study in our laboratory (Hunt, Politzer-Ahles, Gibson, Minai, and Fiorentino, 2013) tested the effect of underinformativeness separately from the influence of lexico-semantic relations. This study used a picture-sentence verification design, in which the truth, falsehood, or underinformativeness of each sentence was based not on real-world knowledge (as was the case for the previous studies) but on the set of items present in a picture presented before the sentence. This made it possible to use identical sentences for all conditions, and manipulate the preceding picture rather than the sentence, thus controlling the lexico-semantic relations within each sentence. This study, like that of Nieuwland and colleagues (2010), found an increased N400 for underinformative critical words, confirming that the upper-bounded interpretation of *some of* was realized online and influenced the access and/or integration of later words in the sentence.

These studies have provided many insights into how scalar implicatures affect online processing as measured by ERPs. However, some open questions remain regarding the time course and neural instantiation of scalar implicature processing. These studies, like the other

N400 studies summarized above, tested whether scalar implicatures can influence the processing of later words in the sentence after the scalar implicature has been computed. As acknowledged by Nieuwland and colleagues, the results of these studies do not "directly reflect full-fledged, online pragmatic inferencing, but rather ... reflect the semantic processing consequences of earlier and relatively implicit pragmatic inferencing" (Nieuwland et al., 2010, p. 341). Because violations in the previous studies only became detectable on words well downstream of the quantifier, these studies cannot make strong claims about how and when the scalar inference is realized. It remains to be seen what pattern of effects may be elicited by processing the scalar implicature itself; this is the question explored in the present study. The three experiments reported in this chapter further investigate scalar implicature processing using a design that dissociates semantic and pragmatic aspects of meaning and examines how each is processed. Importantly, these experiments examine the processing of scalar inferencing at the quantifier itself, rather than at later words in the sentence.

2.1.3. The present studies

The present studies, which were conducted in Mandarin Chinese, adopt a picture-sentence verification design (Wu & Tan, 2009; Tavano, 2010) to compare the neural responses to pragmatically underinformative versus informative sentences that are identical in lexico-semantic content. On each trial a participant is presented with a picture, followed by a sentence that correctly, incorrectly, or underinformatively describes it. Following a picture in which some of the characters are engaging in one activity and others in another (e.g., girls sitting on blankets or on chairs; the upper left portion of Figure 1), a sentence such as "<u>Some of</u> the girls are sitting on blankets" is acceptable, whereas the same sentence is underinformative if it follows a picture in which all of the characters are engaging in the same activity (upper right portion of

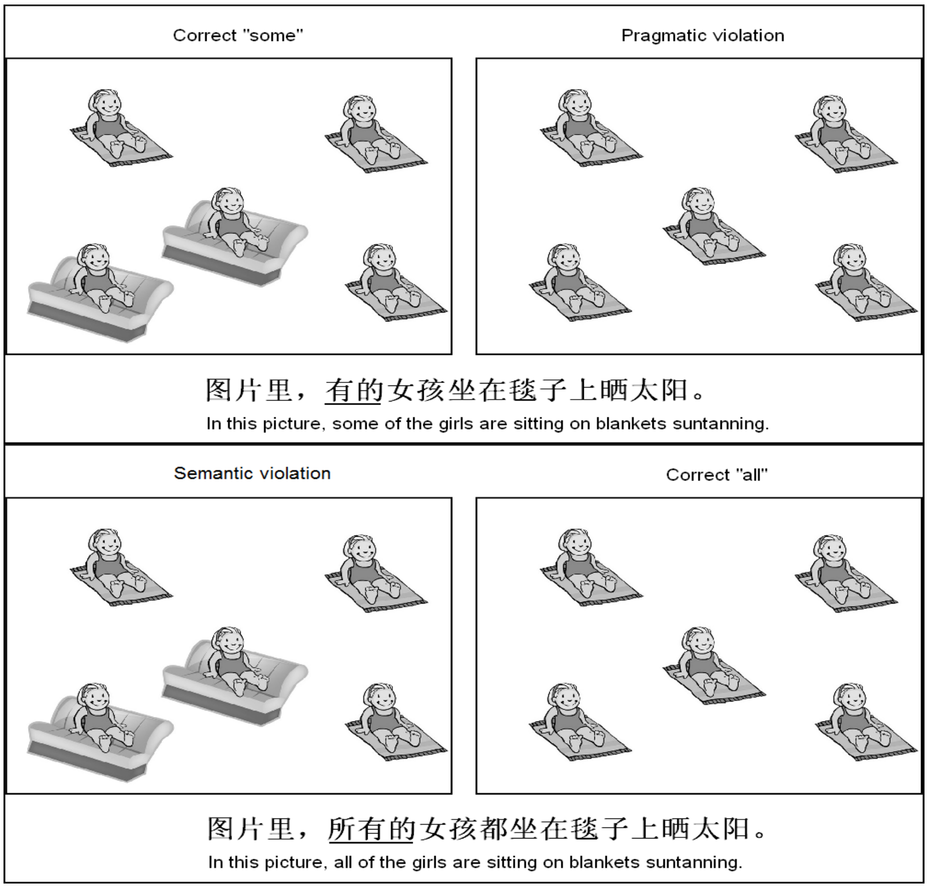

Correct "some"
Pragmatic violation
图片里，有的女孩坐在毯子上晒太阳。
In this picture, some of the girls are sitting on blankets suntanning.
Semantic violation
Correct "all"
图片里，所有的女孩都坐在毯子上晒太阳。
In this picture, all of the girls are sitting on blankets suntanning.

Figure 1. Sample pictures and sentences used in Experiment 1. Upper portion: *some of* sentences preceded by pictures that render them correct (left) or pragmatically incorrect (right). Lower portion: *all of* sentences preceded by pictures that render them semantically incorrect (left) or correct (right).

Figure 1). This design provides a strict control of the context in which the sentence is interpreted, keeping lexico-semantic content identical across conditions. Furthermore, inconsistency becomes detectable at the quantifier itself, making it possible to directly examine the response to underinformative quantifiers rather than the downstream effects of expectations generated by pragmatic inferencing.

The experiments reported here were conducted in Mandarin Chinese, whereas previous online studies of scalar implicature have all used western languages. The characteristics of Mandarin scalar implicature, however, are not different from those of English (see Chi, 2000; Xie, 2003; Tsai, 2004; Rullman & You, 2006; Wu & Tan, 2009). The critical scalar quantifier in the present experiment is *yǒu de* (有的), which is partitive (Xie, 2003; Tsai, 2004) and has a strongly pragmatic interpretation (Wu and Tan's (2009) adult participants reported a pragmatic interpretation of *yǒu de* in 89% of trials). It is roughly equivalent in meaning to the English partitive *some of*, which robustly elicits a pragmatic interpretation (Grodner et al., 2010; Degen & Tanenhaus, 2011).

Experiment 1 tested a factorial manipulation of picture type (in *Some*-type pictures, some characters are engaging in one activity and some in another, whereas in *All*-type pictures all characters are engaging in the same activity) and the quantifier used in the sentence (*some of—yǒu de* 有的—versus *all of—suǒyǒu de* 所有的); see Figure 1 for example pictures and sentences. When used in a sentence following an *All*-type picture, the quantifier *some of* is semantically consistent but pragmatically inconsistent with the picture; when used in a sentence following a *Some*-type picture, the quantifier *all of* is semantically inconsistent with the picture (the inconsistency is due to the inherent semantics of *all*, not due to a pragmatically-enriched meaning).[6] Thus, the experiment has a 2 (Quantifier) × 2 (Consistency) design. Crucially, both inconsistent conditions are compared with lexically matched controls: *some of* following a *Some*-type picture formed the control for the inconsistent *some of* condition, and *all of* following an *All*-type picture formed the control for the inconsistent *all of* condition. In this design, after seeing a picture the participant can form an expectation about the upcoming quantifier—in other words, she can verbally pre-encode the sets as *Some*-type or *All*-type sets (Huang, Hahn, &

[6] Note that, at the position of the quantifier, participants could not be certain whether the inconsistent *all of* sentences were consistent or not with the picture. For instance, if a picture showed some girls sitting on chairs and some sitting on blankets, a sentence beginning "All of…" could be felicitously continued as "All of the girls are wearing hats" or "All of the chairs have girls sitting on them". A similar possibility exists for the *some of* sentences; for instance, a picture showing a group of girls all sitting on chairs could be felicitously continued as "Some of the girls are happy". None of these sentence types was included in the experiment; mismatches between picture and quantifier always led to sentences that were ultimately inconsistent.

Snedeker, 2010; Hartshorne & Snedeker, submitted). Thus, both inconsistent *some of* and inconsistent *all of* are words that are unexpected in their context. Including the *all of* conditions makes it possible to examine the pragmatically inconsistent *some of* condition for effects that are unique to pragmatic processing, above and beyond the effect of seeing an unexpected word.

Experiment 2 tests whether inferential processes involved in comprehending an underinformative sentence interact with lexico-semantic processes, by factorially manipulating the presence or absence of a pragmatic violation early in the sentence with the presence or absence of a lexico-semantic violation on a content word later in the sentence. This is done by using the same picture-sentence verification design as in Experiment 1, and additionally manipulating the lexical consistency between the picture and the sentence: lexically inconsistent sentences have objects (downstream of the quantifier) that do not match any of the objects portrayed in the preceding picture. Thus, Experiment 2 has a 2 (Pragmatic Consistency) × 2 (Lexical Consistency) design, in which sentences are lexically identical across conditions but the pictures preceding the sentences vary.

Experiment 3 is a replication of Experiment 1 using auditory stimulus presentation rather than visual. Furthermore, in Experiment 3 additional measures of participants' pragmatic abilities and sensitivity to scalar implicature were collected, in order to examine potential individual variation in ERP responses to pragmatically inconsistent quantifiers.

2.2. EXPERIMENT ONE

2.2.1. Methods

2.2.1.1. Participants

Data were collected from 23 right-handed Mandarin native speakers (10 females, age range 18-27, mean 20.8) from mainland China who were students at the University of Kansas. Four of these participants were excluded from the statistical analysis because of excessive artifacts in their recordings. All participants had normal or corrected-to-normal vision and were right-handed according to the Edinburgh Handedness Inventory (Oldfield, 1971). All participants provided their informed consent and received payment, and all methods for the study were approved by the Human Subjects Committee of Lawrence at the University of Kansas.

2.2.1.2. Materials[7]

One hundred sixty sets of picture arrays were created for the critical trials (see Figure 1 for an example set). Each picture array included three to five actors or items. In the *All*-type picture array from each set, all of the actors were interacting with identical objects (for instance, four girls were all sitting on blankets, or five baskets were all holding pumpkins). In the *Some*-type picture array from each set, a subset of the actors was interacting with one type of object, and the rest were interacting with a different type of object (for instance, some girls were sitting on blankets and some on sofas, or some baskets were holding pumpkins and some holding bananas). The placement of the actors within the image and the relative locations of actors with different items in the *Some*-type pictures were allowed to vary randomly across sets. All picture arrays were black-and-white cartoons or line drawings, sized 1024 × 768 pixels, and with minimally complex backgrounds. Care was taken to limit pictures to those portraying plausible

[7] A full list of the pictures and sentences used in this and the following ERP experiments (Experiments 2 and 3) is available on request.

events. The base materials for the pictures were taken from freely available clipart from two published databases (Bonin, Peereman, Malardier, Méot, & Chalard., 2003; Szekely et al., 2004) and Google Images, and further edited using Adobe Photoshop, the GNU Image Manipulation Program, and Microsoft Paint by two paid graphic arts students from Peking University and the author.

For each set of picture arrays, *some of* and *all of* sentences were written to match the *All*- and *Some*-type arrays (see Figure 1). Each sentence began with "图片里" ("in this picture"), followed by a subject quantified by either "有的" (*yǒu de*, *some of*), or "所有的" (*suǒyǒu de*, *all of*), followed by a verb and aspect marker, object, and an additional phrase to separate the object from the end of the sentence. Verbs in the critical sentences were marked for progressive, perfective, or prospective aspect. *All of* sentences included the mandatory adverbial 都 *dōu* before the verb (see Li & Thompson, 1981; Jiang et al., 2009). The sentences were written with the help of a paid linguistics student from Peking University who was a native speaker of Mandarin.

Additionally, 148 picture-sentence pairs were created for use as fillers. The filler picture arrays met the same criteria as the critical trials, except that some of them depicted intransitive events. Thirty-seven of these fillers were *Some*-type pictures paired with matching, felicitous *some of* sentences, and thirty-seven were *All*-type pictures paired with matching, correct *all of* sentences. The other seventy-four pictures were paired with sentences that had appropriate quantifiers but either an object that did not match any of the objects in the picture of a verb that did not match the activity shown. Several of these included verbs that yielded semantically anomalous sentences (e.g., "all the scientists are planting squirrels"), whereas most had verbs that were semantically plausible but not congruous with the picture (e.g., "all the boys are going for a walk with their classmates", after a picture in which all the boys are wrestling with their classmates). The filler sentences all included quantifiers that were not used in the critical sentences but were similar in meaning to *all of* or *some of*, or classifier phrases in place of quantifiers. None of the filler sentences used numbers in the place of quantifiers (for discussion of how the presence/absence of numbers and quantifiers in the experimental context may affect the perception of scalar implicature, see Degen, 2009; Grodner et al., 2010; Huang et al., 2010; and references therein). The set of fillers with mismatching pictures and sentences was included to distract participants from the quantifier manipulation in the critical sentences, and the remaining matching fillers were included to maintain a proportion of acceptable sentences of at least 50% during the experiment, assuming that pragmatically infelicitous stimuli are judged as unacceptable.

2.2.1.3. Procedure

Participants were seated in a dimly-lit room about 1 meter in front of a 41-cm CRT monitor. Stimuli were presented at the center of the screen using the Presentation software package (Neurobehavioral Systems). Each trial began with a fixation point presented for 500 ms, followed by a picture which remained on the screen for 4000 ms. The picture was followed by a fixation point of random duration (between 500 and 1500 ms), after which the sentence was presented region by region using the serial visual presentation paradigm. Regions were presented using a variable presentation procedure (see, e.g., Nieuwland et al., 2010), whereby each region was presented at a base duration of 425 ms per region, plus 80 ms for each character more than 3 in the region; because the critical quantifiers were all three characters or less, their presentation durations do not differ across conditions. The interstimulus interval was 400 ms for all regions.[8] Twenty percent of trials were followed by comprehension questions or acceptability judgments (see below), which were presented on the screen for 5000 ms or until the participant's response. Each trial was followed by a blank screen for 1500 ms before the start of the next trial. The experiment was divided into six blocks of approximately 50 sentences each, and participants were given short breaks between the blocks. Participants were instructed not to blink during the presentation of the sentences.

Participants performed a mixture of acceptability judgments and comprehension questions. On ten percent of trials, after the sentence ended, a question that probed information about the picture and was irrelevant to the sentence was presented (e.g., after the sentence "In this picture, some of the girls are sitting on blankets", the comprehension question "Are the girls wearing swimsuits?" appeared). In an additional ten percent of trials, the sentence was followed instead by an acceptability judgment (the question "对不对," "Is that correct?"). Participants were not given explicit instructions about what criteria to consider in judging the sentences, unless they asked for clarification; if they asked, they were instructed to judge, based on their own intuition, whether the sentence was consistent with the picture and described it appropriately. The experimenter stressed that some sentences had no right or wrong answer and that the experiment was meant to measure the participant's own language intuitions. The comprehension questions were included to prevent participants from being able to adopt a strategy of only paying attention to the quantifiers and the number of objects in a picture, and the acceptability

[8] An 800-ms stimulus onset asynchrony (400-ms word presentation, 400-ms interstimulus interval) has been found to be natural and comfortable for Chinese readers in previous studies (e.g., Jiang et al., 2009), but the regions used in the present study tended to be longer than the regions used in those studies, and pilot participants reported the variable presentation rate described above to be the most comfortable.

judgments were included to ensure that participants pay attention to the sentence rather than just try to remember the picture. Acceptability judgment prompts were allotted to six of the forty pragmatically infelicitous sentences for each participant, making it possible to determine whether participants accepted or rejected these sentences when making an explicit judgment. Participants responded to both the comprehension questions and acceptability judgment prompts using the left and right buttons on a mouse.

The experimental sentences were divided into four lists according to a Latin square design, such that every sentence appeared once in each condition across lists but no sentence or picture was repeated within a list. The item order in the list was fully randomized for each participant. The first block of the experiment was preceded by a practice block of seven trials which followed the same presentation procedure as the main experiment but did not include any quantifier-related violations. The practice sentences included some sentences with existential quantifiers (e.g., "图片里有。。。," "in the picture there are") and some without quantifiers (e.g. "图片里的小狗," "the dogs in the picture are…"). Feedback was given for behavioral responses in the practice block, but not in the main experiment. The recording itself took 70 to 80 minutes.

2.2.1.4. Data acquisition and analysis

The EEG was continuously recorded using an elastic electrode cap (Electro-Cap International, Inc.) containing 32 Ag/AgCl scalp electrodes organized in a modified 10-20 layout (midline: FPZ, FZ, FCZ, CZ, CPZ, PZ, OZ; lateral: FP1/2, F7/8, F3/4, FT7/8, FC3/4, T3/4, C3/4, TP7/8, CP3/4, T5/6, P3/4, O1/2). Polygraphic electrodes were placed at the left and right outer canthi for monitoring horizontal eye movements, above and below each eye for monitoring blinks, and on the left and right mastoids. The left mastoid served as a reference during data acquisition and AFz served as the ground. Impedances for scalp electrodes and mastoids were kept below 5 kΩ. The recordings were amplified by a Neuroscan Synamps2 amplifier (Compumedics Neuroscan, Inc.) with a bandpass of 0.01 to 200 Hz, and digitized at a sampling rate of 1000 Hz.

The continuous EEG was re-referenced to the average of both mastoids and segmented into epochs from 1000 ms before to 2000 ms after the presentation of the critical word. Based on visual inspection, trials containing excessive muscle artifact or alpha activity within the epoch of 200 ms before to 1200 ms after the onset of the stimulus were excluded from the analysis. Following artifact rejection, the data were demeaned using the mean amplitude of each epoch (Groppe et al., 2009), and an independent components (ICA) decomposition algorithm (Makeig et al., 1996) was applied to remove ocular artifacts. After artifact correction, the EEG was visually inspected again to remove trials in which any artifact remained. A total of 18.8% of trials was rejected in this way (18.9% of pragmatically inconsistent *some of* trials; 16.2% of correct *some of* trials; 20% of semantically inconsistent *all of* trials; and 20.1% of consistent *all of* trials); a repeated measures ANOVA revealed that marginally more *some of* than *all of* trials were kept in the analysis ($F(1,18) = 3.49$, $p = .078$) and that there was no significant effect of consistency or interaction between quantifier or consistency in terms of trials kept (*p*s > .16). Participants with fewer than 25 trials remaining for any condition after artifact rejection were excluded from the analysis. Subsequently, data epochs were baseline-corrected using a 200-ms pre-stimulus baseline and averaged to calculate ERPs.

Time windows for analysis were chosen based on visual inspection of the data, and mean ERP voltage amplitudes were compared using repeated measures ANOVAs involving the factors Consistency (consistent, inconsistent), Quantifier (*some of*, *all of*), and the topographical factor Region. Midline and lateral regions were analyzed separately. For the lateral ANOVA, regions were defined by averaging within the following electrode groups: left anterior (F7, F3, FC3), left central (T3, C3, CP3), left posterior (T5, P3, O1), right anterior (F4, F8, FC4), right central (C4, T4, CP4), and right posterior (P4, T6, OZ). For the midline ANOVA, regions were defined as follows: anterior (FZ, FCZ), central (CZ, CPZ), and posterior (PZ, OZ). The Huynh-Feldt correction was applied to *F*-tests with more than one degree of freedom in the numerator.

2.2.2. Results

2.2.2.1. Behavioral results

Participants responded both to comprehension questions irrelevant to the interpretation of the quantifier and to acceptability judgment prompts during the course of the experiment. Behavioral data from one participant were lost due to a data logging error, leaving eighteen participants for the behavioral data analysis. In the comprehension task, mean accuracy rates were 86.1% for the pragmatically infelicitous condition (*some of* sentence following an *All*-type picture), 77.5% for consistent "some", 82.8% for semantically inconsistent (*all of* sentence following a *Some*-type picture), and 78.2% for consistent "all". A repeated measures ANOVA revealed no significant differences in mean accuracy across conditions ($F(3, 51) < 1$).

Acceptability judgments on the pragmatically underinformative sentences have no correct or incorrect answer, given that participants can interpret such sentences semantically or pragmatically. Across participants, 39.8% of pragmatically underinformative sentences were judged as correct, indicating a semantic judgment; in comparison, only 19.6% of semantically inconsistent sentences were judged as correct. The difference was significant by participants ($t(17) = -4.47$, $p < .001$), indicating that participants accepted pragmatically infelicitous sentences more often than semantically inconsistent sentences. As for the remaining conditions, which do have clear expected judgments, mean accuracy rates were 78.7% for the consistent "some" condition, 80.4% for the semantically inconsistent condition, and 85.5% for the consistent "all" condition. A repeated measures ANOVA revealed no significant differences across conditions ($F(2, 34) < 1$).

Several previous studies have distinguished between pragmatic and semantic responders (Noveck & Posada, 2003; Bott & Noveck, 2004; Tavano, 2010; Hunt et al., 2013). Thus, participants were divided into groups using the following criteria: participants who made 5 or more semantic responses (1 or fewer pragmatic responses—each participant judged 6 underinformative sentences, see section 4.1.3) to the underinformative trials were classified as semantic responders, those who made 5 or more pragmatic responses (1 or fewer semantic responses) were classified as pragmatic responders, and those who made 2 to 4 semantic responses (no more than 4 responses of a given type) were classified as inconsistent responders. Five participants met the criteria to be considered semantic responders, while two were pragmatic responders and eleven inconsistent; there were not enough consistent responders to form participant groups for the ERP analysis.[9] There was a greater number of inconsistent responders in the present study than in some previous studies (Noveck & Posada, 2003; Tavano, 2010), which is consistent with Feeney and colleagues (2004), who found that participants tended to respond inconsistently to underinformative sentences when the variety of stimulus conditions is large (see Section 2.2.1.2 for more information about the conditions included in the present experiment).

[9] Using slightly more lax criteria (4 or more semantic responses for semantic responders, 4 or more pragmatic responses [2 or fewer semantic responses] for pragmatic responders, and 3 semantic [3 pragmatic] responses for inconsistent responders), 8 responders were classified as semantic responders, 3 as pragmatic, and 7 as inconsistent.

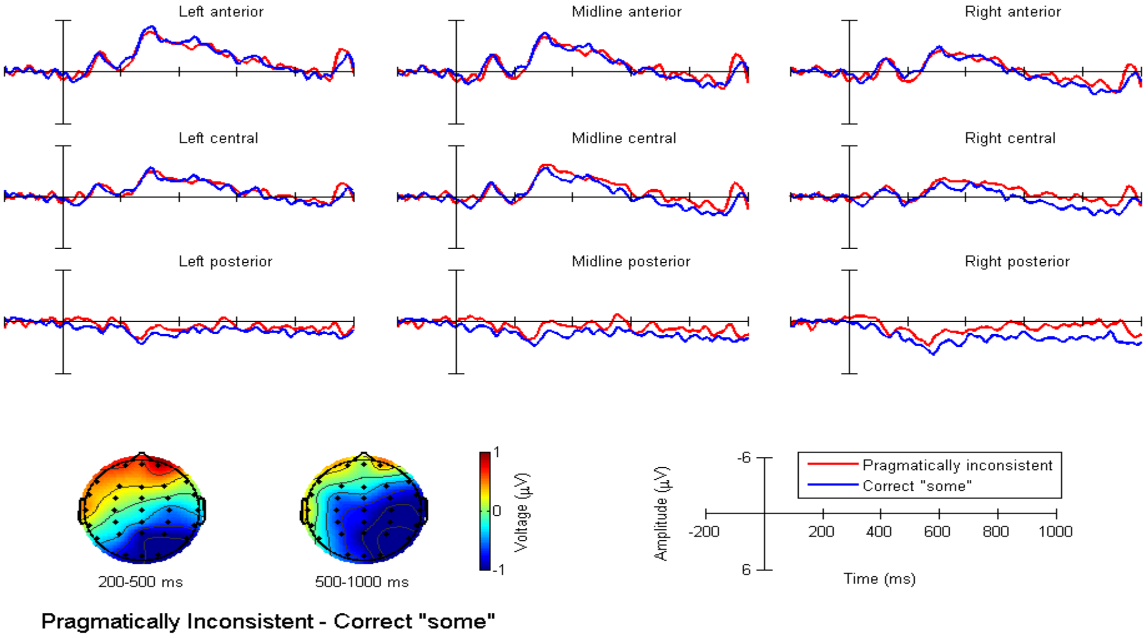


Figure 2. Effect of pragmatic inconsistency in Experiment 1. Upper portion: Grand average ERPs (a 30 Hz low-pass filter was applied for plotting) at nine scalp regions. Lower portion: Topographic maps formed by subtracting the correct *some of* condition from the pragmatically incorrect condition over two time windows**.**

2.2.2.2. ERP results

Visual inspection of the waveforms (Figure 2 and Figure 3) suggests that semantically inconsistent *all of* elicited a less negative ERP than consistent *all of* from about 200 to 500 ms in the anterior and central regions, whereas pragmatically inconsistent *some of* elicited a sustained negative ERP compared to consistent *some of* in the right posterior regions. Thus, ANOVAs were conducted on the mean ERP amplitudes for the 200-500 ms and 500-1000 ms time windows; the omnibus ANOVA results are shown in Table 1.

2.2.2.2.1. 200-500 ms

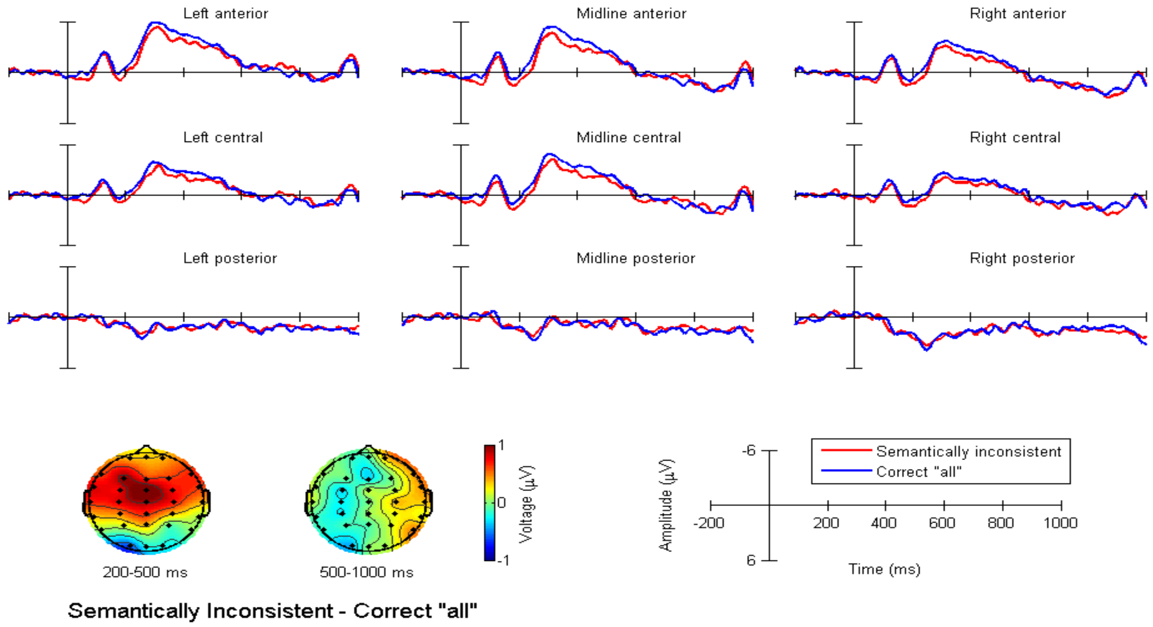


Figure 3. Effect of semantic inconsistency in Experiment 1. Upper portion: Grand average ERPs (a 30 Hz low-pass filter was applied for plotting) at nine scalp regions. Lower portion: Topographic maps formed by subtracting the correct *all of* condition from the semantically inconsistent condition over two time windows.

The ANOVA revealed a significant interaction of Consistency and Region.[10] The interaction was resolved by testing the effect of Consistency at each region. Both types of inconsistent quantifier elicited significantly more positive ERPs than controls in the left anterior region ($F(1,18) = 4.52$, $p = .048$), marginally more positive in the midline anterior ($F(1,18) = 3.91$, $p = .063$) and left central ($F(1,18) = 3.21$, $p = .090$) regions, and marginally more negative ERPs in the right posterior region ($F(1,18) = 4.08$, $p = .059$); the simple effect of consistency did not reach significance in any other region (*p*s > .143).[11]

[10] There were also effects of Quantifier by Region in this time window and of Quantifier in the later time window. These, however, are not of theoretical interest since they involve direct comparison between different words, and thus are not discussed here. The significant main effects of Region are also not discussed since they do not reveal differences based on the experimental manipulation.

[11] Visual inspection of the waveforms and topographic plots (Figure 2 and Figure 3) suggests that the posterior negativity revealed in the Consistency by Region interaction was due to the pragmatically inconsistent quantifiers, whereas the anterior positivity was present in both conditions—i.e., that semantically inconsistent quantifiers

Effect	200-500 ms	500-1000 ms
Quantifier	*F*(1,18) = 1.07 *F*(1,18) = 2.42	*F*(1,18) = 4.04* *F*(1,18) = 1.07
Consistency	*F*(1,18) = 0.15 *F*(1,18) = 0.18	*F*(1,18) = 2.08 *F*(1,18) = 2.34
Region	*F*(5,90) = 49.19**** *F*(2,36) = 38.60****	*F*(5,90) = 20.67**** *F*(2,36) = 11.12***
Quantifier × Consistency	*F*(1,18) = 1.92 *F*(1,18) = 2.46	*F*(1,18) = 2.63 *F*(1,18) = 1.04
Quantifier × Region	*F*(5,90) = 2.98** *F*(2,36) = 1.90	*F*(5,90) = 0.05 *F*(2,36) = 0.48
Consistency × Region	*F*(5,90) = 6.73*** *F*(2,36) = 7.25***	*F*(5,90) = 0.65 *F*(2,36) = 0.64
Quantifier × Consistency × Region	*F*(5,90) = 0.31 *F*(2,36) = 0.14	*F*(5,90) = 3.06** *F*(2,35) = 0.50

Table 1. Results of the lateral and midline omnibus ANOVAs in Experiment 1 at two time windows, with each cell showing the lateral ANOVA result above and the midline ANOVA result below. *.05 < p < .1; ***p* < .05; ****p* < .005; *****p* < .001.

2.2.2.2.2. 500-1000 ms

In the later time window there was a significant interaction of Quantifier, Consistency, and Region in the lateral ANOVA only. Resolving the interaction by Quantifier revealed that pragmatically inconsistent quantifiers elicited both a significant main effect of Consistency (*F*(1,18) = 4.56, *p* = .047) and a Consistency by Region interaction (*F*(5,90) = 3.07, *p* = .039), but neither an interaction nor a main effect of Consistency was observed for the semantically

elicited an anterior positivity only, whereas pragmatically inconsistent quantifiers elicited both an anterior positivity and posterior negativity. However, the interaction of Quantifier and Consistency in the omnibus ANOVA did not reach significance (see Table 1), providing no evidence for differential ERP responses to semantic and pragmatic inconsistencies in this time window.

inconsistent quantifiers (*F*s < 1). For the pragmatically inconsistent quantifiers, the main effect of Consistency was due to a more negative ERP for inconsistent than consistent quantifiers in this time window, and the interaction with Region was due to the fact that the simple effect of Consistency for *some of* reached significance at the right central (*F*(1,18) = 7.09, *p* = .016) and right posterior (*F*(1,18) = 11.63, *p* = .003) regions, but not at other regions (*p*s > .108).

2.2.3. Discussion

This experiment tested whether the pragmatic meaning of a scalar quantifier affects processing immediately when the quantifier itself is read, and how the detection of pragmatic implicature violations is manifested electrophysiologically when lexico-semantic differences are controlled for. Both quantifiers that were semantically inconsistent with a context and those that were pragmatically inconsistent elicited a less negative anterior ERP than controls in an earlier (200-500 ms) time window. This early effect indicates that the pragmatic interpretation of the scalar quantifier was used rapidly during processing, since the quantifier was only inconsistent with its context when interpreted pragmatically; this effect was not unique to scalar implicature processing, however, as it was also elicited by unexpected, semantically inconsistent quantifiers. Effects unique to scalar implicature processing were observed later in the epoch (500-1000 ms), at which time pragmatically inconsistent but not semantically inconsistent quantifiers elicited a sustained posterior negativity. While this negativity also appeared earlier in the epoch with a topography similar to an N400 effect, it is apparent from the waveforms that the effect is more likely the beginning of a sustained negativity lasting throughout the epoch; note that Pijnacker and colleagues (2011) also found a dissociation between a transient N400 elicited by lexico-semantic violations, and a more long-lasting negativity elicited by discourse processing.

In experimental contexts like this one, rapid effects of pragmatic inconsistency could be due to participants' ability to verbally pre-encode the picture contexts as *Some*-type or *All*-type contexts, and then make a forward prediction about the quantifier that will appear in the sentence (Huang et al., 2010; Hartshorne & Snedeker, submitted). Indeed, the presence of an early effect is not surprising, as previous research has already shown that pragmatic expectations about upcoming words can modulate ERPs as early as the N400 (Van Berkum, 2009; Nieuwland et al., 2010; Hunt et al., 2013). However, it is unlikely that the results of the present experiment are due only to effects of seeing an unexpected word. First of all, unexpected linguistic input typically elicits a N400 or P300/P600 effect (Lau et al., 2008; Bornkessel-Schlesewsky et al., 2011), whereas the topography and polarity of the early effect in the present experiment was consistent with neither of these. Rather, the effect is consistent in timing and topography with the Nref, a negativity suggested associated with establishing the antecedent of a word (Van Berkum et al., 2007). In the present experiment, the smaller negativity for inconsistent quantifiers may reflect a decrease in effort made to link *all of* or *some of* with an antecedent when the participant recognizes it to be pragmatically or semantically inconsistent with the context. More importantly, if participants were making predictions based on verbal pre-encoding, then *all of* and *some of* would both be unexpected; nevertheless *some of* elicited a qualitatively different effect later in the epoch.

In the present experiment, pragmatically inconsistent *some of*, but not semantically inconsistent *all of*, elicited a sustained negativity in the late time window. The present experiment is not the first to observe such an effect. Sustained negativities have also been observed on sentences in which readers must re-compute a discourse model or revise a discursive inference (Baggio et al., 2008; Pijnacker et al., 2011). The former study examined the Dutch equivalents of sentences such as "The girl was writing a letter when her friend <u>spilled</u> coffee on the paper" (note that the verb is sentence-final in the Dutch equivalent of the sentence, which was used in the experiment), in which the main clause allows the reader to infer that the girl will eventually finish writing the letter but following clause cancels that inference. The authors observed a sustained anterior negativity on the sentence-final verb, which they argue reflects the process of re-computing a previously computed discourse model in which the girl finishes writing the letter. The experiment by Pijnacker and colleagues (2011) presented participants with *modus ponens* reasoning problems in which a normally logical conclusion is defeated by contextual information: for example, the conclusion in problem (4) would normally be logical, but is disabled by the context presented in (4a).

4) a. **Context**: Lisa probably lost a context lens.
   b. **Premise 1**: If Lisa is going to play hockey, then she will wear contact lenses.
   c. **Premise 2**: Lisa is going to play hockey.
   d. **Conclusion**: Lisa will wear contact lenses.

They found that the final words of the disabled conclusions elicited a broad sustained negativity relative to controls, and they interpreted this effect as representing the revision of the discursive inference that would normally lead from the premises to the conclusion. The sustained negativity observed in the present study has a similar morphology and latency as the negativities observed in those experiments. It may be the case that the negativity observed in the present study reflects revision of the reader's interpretation of the quantifier's meaning (inhibition of the pragmatic reading and retrieval of the semantic reading) after the reader realizes that the pragmatic reading is inappropriate.

While psycholinguistic models assert that realizing and/or cancelling a pragmatic inference may involve processing costs (Katsos & Cummins, 2010, Hartshorne & Snedeker, submitted; see also Garrett & Harnish, 2007), they do not yet articulate precisely what sort of costs or mechanisms this computation entails (see Bott et al., 2012, for further discussion). Thus, the next experiment examines whether the canceling or suppression of a pragmatic inference (reflected by the sustained negativity in Experiment 1) interacts with lexico-semantic processing. This experiment factorially manipulates the presence of a lexico-semantic violation (i.e., a sentence object that does not match the objects in the picture) and the felicity of the quantifier *some of* upstream of the violation; example pictures and sentences for Experiment 2 are shown in Figure 4. For example, the sentence “Some of the girls are sitting on blankets suntanning” is pragmatically and lexically correct when preceded by a sentence in which some girls are sitting on blankets and some sitting instead on couches (depicted in the upper-left portion of Figure 4). The same sentence is pragmatically correct but lexically incorrect when none of the girls are sitting on blankets but not all the girls are sitting on the same thing (upper-right portion). The sentence is pragmatically incorrect but lexically correct when in fact all of the girls are sitting on blankets (lower-left portion). Finally, when all the girls are sitting on the same thing and that thing is not a blanket, the sentence is both pragmatically and lexico-semantically incorrect (lower-right portion), making it possible to examine how the neural response to lexico-semantic inconsistency at the object position interacts with the processing of the pragmatic inconsistency previously instantiated at the quantifier position.

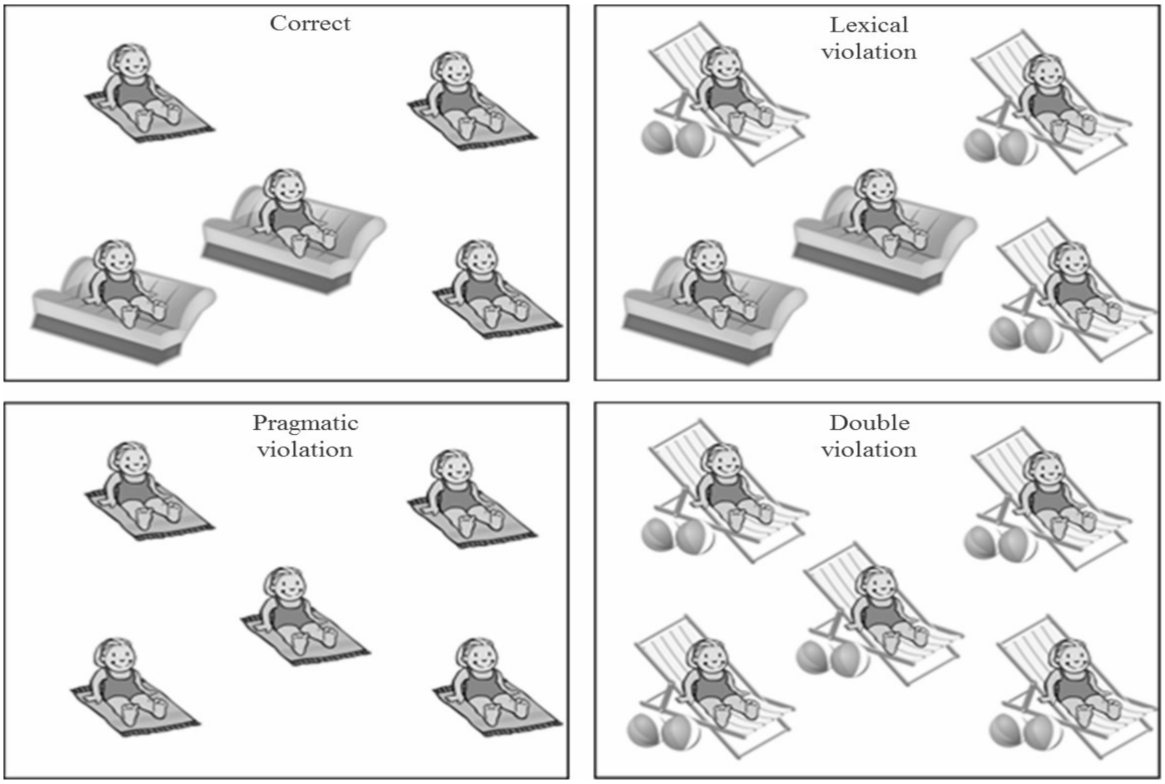


Figure 4. Sample pictures used in Experiment 2; in this sample, all pictures were followed by the sentence 图片里，有的女孩坐在毯子上晒太阳 ("In the picture, some of the girls are sitting on blankets suntanning"). In a given trial, only one of the pictures was shown before the sentence. The condition labels on the picture are for expository purposes only and were not included in the experiment.

Lexico-semantic picture-sentence mismatches have been shown to elicit robust N400s (Knoerfle, Urbach, & Kutas, 2011). If the ongoing pragmatic revision process after encountering an infelicitous quantifier affects lexico-semantic processing, either by limiting the extent to which the parser commits to predictions about upcoming material or by using the same processing resources that would otherwise be used for lexico-semantic prediction and integration, then the N400 effect for lexico-semantic violations at the object position should be modulated. For instance, Panizza and colleagues (2011) found that participants in a visual world eye-tracking experiment were slower to look to an unambiguous target (e.g., slower to look to a referent with paper clips after *paper clips* had already been heard) when the target word was in an upward entailing environment, which supports scalar implicature (e.g., "A boy has some of the paper clips; click on him"), than when it was in a downward entailing environment, which does not (e.g., "If a boy has all of the paper clips, then click on him"). The authors suggest that generating a scalar implicature may have interfered with participants' ability to use disambiguating phonological information for lexico-semantic integration. In a similar vein, Experiment 2 of the current study tests whether revising an underinformative scalar inference interferes with lexico-semantic integration between the picture and the sentential object. The Quantifier by Consistency manipulation at the quantifier position from Experiment 1 was also included in this experiment, in order to test whether the effect obtained in that experiment would be replicated. (The pragmatically inconsistent "some" and correct "some" conditions were included in the critical items; items corresponding to the semantically inconsistent "all" and correct "all" of Experiment 1 were included in the fillers for this experiment.) While the primary motivation for Experiment 2 was to examine the interaction of pragmatic and lexical processing rather than effects of modality, auditory presentation of sentences was found both to be comfortable for participants and to reduce the duration of each trial. For this reason, sentence stimuli were presented auditorily rather than visually in Experiment 2.

2.3. EXPERIMENT TWO

2.3.1. Methods

2.3.1.1. Participants

Twenty-three Peking University students (9 females; mean age 22.5 years, range 18-26) who were native speakers of Mandarin participated in the study. Three were excluded from the statistical analysis due to excessive artifacts in their recordings, leaving a total of 20 participants in the final analysis. All participants had normal or corrected-to-normal vision and were right-handed according to the Chinese Handedness Survey (Li, 1983). All participants provided their informed consent and received payment, and all methods for the study were approved by the Ethics Committee of the Department of Psychology, Peking University, and the Human Subjects Committee of Lawrence at the University of Kansas.

2.3.1.2. Materials

Two hundred and sixty sets of picture arrays were designed according to the same criteria as in Experiment 1. Each *Some*- and *All*-type picture array had two versions, such that in the first version the object being interacted with by some or all of the characters matched the object mentioned in the associated sentence, and in the second version it mismatched. At the object position, this formed a 2 (Lexical Consistency) × 2 (Pragmatic Consistency) design: sentences with correct objects, sentences with lexical mismatches at the object position, sentences with correct objects but a pragmatic violation upstream, and sentences with both a pragmatic violation upstream and a lexically incorrect object. It formed a one-factor design at the quantifier position: sentences with consistent quantifiers and those with pragmatically inconsistent quantifiers (each of these conditions collapsed across lexically consistent and inconsistent sentences, since at the quantifier position the lexical mismatch has not yet been encountered). A sample stimulus set is shown in Figure 4. Critical sentences were written so that none of the critical objects were at the end of the sentence. All the critical objects used were either 2 or 3 syllables long.[12]

[12] The 200 plausible most plausible all-type pictures were normed with a sentence completion task to select pictures in which the objects were most identifiable. Twenty-eight students from Beijing Union University participated in the task. Participants were presented with the pictures along with sentence fragments up to but not including the objects (e.g. "图片里，所有的女孩都坐在。。。", "In the picture, all the girls are sitting on…") and asked to complete the sentence. For critical stimuli for the ERP experiment, the 160 sentence-picture pairs whose objects had the highest

Two hundred forty filler sentences were prepared, using picture-sentence pairs that had not been chosen for the critical items as well as new picture-sentence pairs. Eighty were used to test the semantic violation at quantifier position (forty correct *all* and forty semantically inconsistent *all of* sentences, counterbalanced across participants); these sentences, together with the critical sentences, making it possible to test whether the Consistency by Quantifier interaction reported in Experiment 1 could be replicated. Of the remaining fillers, eighty were correct *all of* sentences that were not analyzed, and the last 80 were sentences using other quantifiers. Of those 80, 40 used *some*-like quantifiers (e.g. 有一些 *a few*) and 40 used *all*-like quantifiers (e.g. 每个 *every*). None included quantifier-related violations; 40 were entirely correct, 20 mismatched with the picture at the object position, and 20 mismatched at the verb position. (Out of each of these types, half of the items used *all*-like quantifiers and half used *some*-like.)

Auditory stimuli were read by a female native speaker from the Peking University Chinese department, who was instructed to avoid placing contrastive stress on the quantifiers and objects. The recordings were digitized at 22050 Hz using CoolEdit Pro (Syntrillium Software) and segmented using Praat (Boersma & Weenik, 2012), and the onset latencies of the quantifiers and objects were measured. The onset of the quantifier *yǒu de* (*some of*) was defined as the point of lowest intensity between the preceding syllable *lǐ* and the *yǒu*, which in most tokens also coincided with a perceptible change in phoneme quality and preceded, by 10-20 ms, a 200-400

cloze probability were chosen, with the condition that a pair was not chosen if any identical objects were given in response to both pictures. All sentences chosen had an object cloze probability above 46% (mean 81%). Due to reorganization of target and filler stimuli to avoid repetition of target objects, two picture/sentence pairs that had not been cloze tested were later introduced into the critical stimuli.

Hz drop in frequency of the second through fourth formants. The onset of the quantifier *suǒyǒu de* (*all of*) was defined as the onset of high-frequency energy in the spectrogram. Onsets of objects were measured as the audible onset of the first consonant of the word (plosives were measured at the burst), except in two cases where the onset of the first consonant of the second syllable was measured since this was the point of disambiguation for the critical word. The latency between quantifier onsets and object onsets in the critical sentences was 1309 ms on average (SD = 203 ms, range 832-2127 ms).

The 400 trials (160 critical *some of* sentences, 80 *all of* fillers, and 160 other fillers) were arranged into four lists in a Latin square design. Each list contained 40 trials per object condition. For the *all of* sentences tested, each list contained 40 trials per condition (correct "all", semantically inconsistent "all").

Each list was divided into five blocks of 80 trials each, such that the first trial in each block was a filler sentence. Each block was pseudorandomized according to the following criteria: no more than three trials of the same condition could appear consecutively, no more than four correct or incorrect trials could appear consecutively, no more than six *Some*-type or *All*-type pictures could appear consecutively, and no more than six *some of* or *all of* sentences could appear consecutively. The order of trials was kept the same for each list, such that a given item appeared at the same position (but in different conditions) in every list, and each of the lists adhered to the above constraints.

2.3.1.3. Procedure

Participants were seated comfortably in a dimly lit and electromagnetically shielded room, about 80 cm in front of a 51-cm CRT monitor. Pictures were presented on the monitor and sentences were presented through tube earphones (Etymotic Research, Inc.). Stimulus presentation and recording of behavioral responses was controlled using Presentation software (Neurobehavioral Systems). Each trial began with a fixation point presented at the center of the screen for 500 ms, followed by the picture, which was presented at the center of the screen for 4000 ms. After this time the picture disappeared and was immediately replaced by a fixation point at the center of the screen, which remained on the screen throughout the presentation of the auditory sentence. The sentence began between 500 and 1500 ms after the appearance of the fixation point. After the end of the sentence, a 1-7 scale appeared on the screen. The extremes (1 and 7) were labeled "一致" ("consistent") and "不一致" ("inconsistent"); the sides of the scale on which these extremes appeared were counterbalanced across participants.

The participants' task was to rate how consistent the sentence was with the preceding picture within 3000 ms. The rating task was chosen to encourage participants to pay attention to the entire sentence and thus reduce the possibility that they could complete the task strategically simply by matching numbers of items in the picture with quantifiers in the sentence; rating tasks have been used in previous online studies investigating quantification (Urbach & Kutas, 2010) and scalar implicature (Foppolo, 2007). After the rating task was complete, the trial was followed by a 2500 ms blank screen before the fixation point signaling the beginning of the next trial.

After every 80 trials the participants were given a break. In addition, after every 20 trials they were given a 15-second break, during which time a message appeared on the screen asking them to relax briefly. The formal experiment was preceded by a practice session consisting of 10 trials. The trial structure and picture formats were identical to those used in the main experiment, but no violations involving picture-object mismatch or pragmatic underinformativeness were included. The recording took about 100 minutes.

2.3.1.4. Data acquisition and analysis

The EEG was continuously recorded using an elastic electrode cap (Brain Products, Munchen, Germany) containing 64 tin electrodes organized according to the 10-20 system. Additional channels were placed above the right eye and at the outer canthus of the left eye for monitoring vertical and horizontal electro-oculograms (EOGs), respectively. An electrode placed on the tip of the nose served as the reference during data acquisition, and AFz served as the ground. Impedances were kept below 10 kΩ. The recordings were amplified using a Brain Products Brainamp amplifier with a bandpass from 0.016 to 100Hz, and digitized at a sampling rate of 500 Hz.

The raw EEG was segmented into epochs from 1000 ms before to 4250 ms after the quantifiers (this epoch ensured at least 2000 ms after each critical object). Data were then demeaned using the mean amplitude of each epoch (Groppe et al., 2009), decomposed with an ICA algorithm (Makeig et al., 1996) to remove ocular artifacts, and re-segmented into two separate datasets (one consisting of -200 to 1000 ms epochs time-locked to the quantifiers, and one consisting of -200 to 1000 ms epochs time-locked to the objects). Artifact rejection was performed separately for the quantifier and object data, and ERPs time-locked to the object used a 100-ms post-stimulus baseline rather than a 200-ms pre-stimulus baseline, since the pre-stimulus interval contained sustained effects of processing violations at the quantifier. 11.7% of trials were rejected (9.8% of epochs time-locked to the objects, and 13% of epochs time-locked to the quantifiers); all subjects included in the analysis had at least 29 trials per condition in the object analysis and 25 per condition in the quantifier analysis. The proportion of trials rejected did not differ between conditions in either analysis (objects: *F*s < 1; quantifiers: *F*s < 1.06, *p*s > .315).

The following electrode regions were defined on this cap: left anterior (F1, F3, F5, FC1, FC3, FC5), right anterior (F2, F4, F6, FC2, FC4, FC6), left central (C1, C3, C5, CP1, CP3, CP5), right central (C2, C4, C6, CP2, CP4, CP6), left posterior (P1, P3, P5, PO3, PO7, O1), right posterior (P2, P4, P6, PO4, PO8, O2), midline anterior (Fz, FPz), midline central (Cz, CPz), midline posterior (POz, Oz). For the quantifier position, the analysis used the factors Consistency (consistent, inconsistent), Quantifier (*some of*, *all of*), and Region (6 levels for the lateral ANOVA, 3 for the midline ANOVA). For the object position, the factors were Pragmatic Consistency (consistent, inconsistent), Lexical Consistency (consistent, inconsistent), and Region. The Huynh-Feldt correction was applied to *F*-tests with more than one degree of freedom in the numerator.

2.3.2. Results

2.3.2.1. Behavioral results

The participants' task was to rate the consistency between the picture and the sentence using a 7-point scale. Average ratings are shown in Figure 5. A repeated measures ANOVA on the four critical conditions (correct *some of*, pragmatic violation, lexical mismatch, and double violation) revealed significant effects of Pragmatic Consistency ($F(1,18) = 29.11$, $p < .001$) and of Lexical Consistency ($F(1,18) = 206.68$, $p < .001$), but no significant interaction ($F(1,18) = .03$, $p = .862$). Furthermore, pairwise *t*-tests between all six conditions, with the two-tailed alpha level Bonferroni-corrected to $\alpha = .003$, revealed significant differences for every comparison except correct *some of* vs. correct *all of* ($p > .5$) and the double violation vs. semantically incorrect *all of* ($p = .32$). These results indicate that participants treated correct sentences, pragmatic violations, lexical mismatches, and double violations as decreasingly acceptable, but they did not

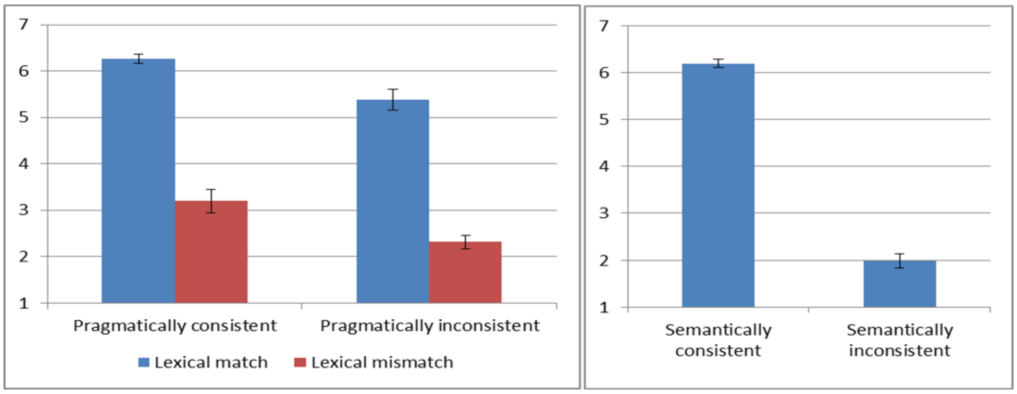


Figure 5. Consistency ratings in Experiment 2 (1=very inconsistent, 7=very consistent) for the *some of* sentences (left) and *all of* sentences (right). Error bars represent ±1 standard error of the mean.

differentiate between the two correct conditions or between double violations (with both pragmatic violation and picture-sentence mismatch) and semantically incorrect "all". Because the present experiment used a gradient rating task rather than a categorical judgment task, it was not possible to classify participants as pragmatic or semantic responders using the same criteria as in Experiment 1 or previous studies (Noveck & Posada, 2003; Bott & Noveck, 2004; Feeney et al., 2004; Hunt et al., 2013).[13]

2.3.2.2. ERP results

[13] Nevertheless, the number of pragmatic responders was assessed using one-tailed independent samples *t*-tests for each participant comparing ratings for pragmatic violations against ratings for correct sentences. Twelve participants reliably rated pragmatically inconsistent sentences lower than correct sentences (*ps* < .05), whereas eight did not. The former group may be considered pragmatic responders (those who interpreted *some* as meaning *not all*), whereas the latter group may be either semantic responders (those who interpreted *some* as meaning *at least one*) or inconsistent responders. Compared to the acceptability judgment task used in Experiment 1, in which most participants were inconsistent, the consistency rating in Experiment 2 yielded a greater number of pragmatic responders.

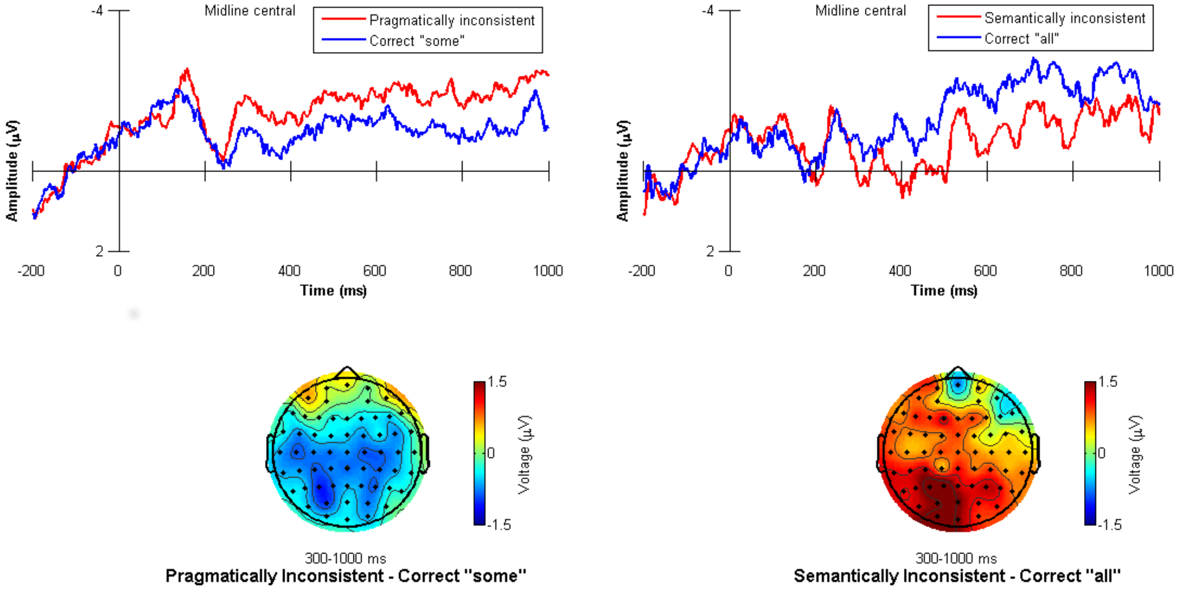


Figure 6. Effects of pragmatic and semantic inconsistency at the quantifier in Experiment 2. Upper portion: Grand average ERPs at the midline central region. Lower portion: Topographic maps formed by subtracting the correct quantifier condition from the corresponding inconsistent quantifier conditions.

The waveforms time-locked to the quantifier position (Figure 6) show a sustained negativity for pragmatically inconsistent quantifiers, similar to the one obtained in Experiment 1 but broader in distribution, and a sustained positivity for semantically inconsistent quantifiers. At the object position (Figure 7), both picture-sentence mismatches and double violations elicited broadly-distributed negativities from about 200-600 ms, whereas both types of objects following pragmatically inconsistent quantifiers elicited a sustained negativity from about 400-1000 ms. In this time window the sustained negativity appeared to be present for the objects following pragmatic violations and for the double violations, but not for the picture-sentence mismatches.

These patterns of effects are examined statistically below; the omnibus ANOVA results for the quantifier and object positions are presented in Table 2 and Table 3, respectively.

Experiment 2 – quantifiers

Effect	**300-1000 ms**
Quantifier	*F*(1,19) = 0.06 *F*(1,19) = 0.08
Consistency	*F*(1,19) = 0.85 *F*(1,19) = 1.70
Region	*F*(5,95) = 71.96**** *F*(2,38) = 34.36****
Quantifier × Consistency	*F*(1,19) = 10.92** *F*(1,19) = 6.10**
Quantifier × Region	*F*(5,95) = 1.30 *F*(2,38) = 0.75
Consistency × Region	*F*(5,95) = 1.51 *F*(2,38) = 0.98
Quantifier × Consistency × Region	*F*(5,95) = 2.83** *F*(2,38) = 0.82

Table 2. Results of the lateral and midline omnibus ANOVAs at the quantifier position in Experiment 2, with each cell showing the lateral ANOVA result above and the midline ANOVA result below. *.05 < p < .1; ***p* < .05; ****p* < .005; *****p* < .001.

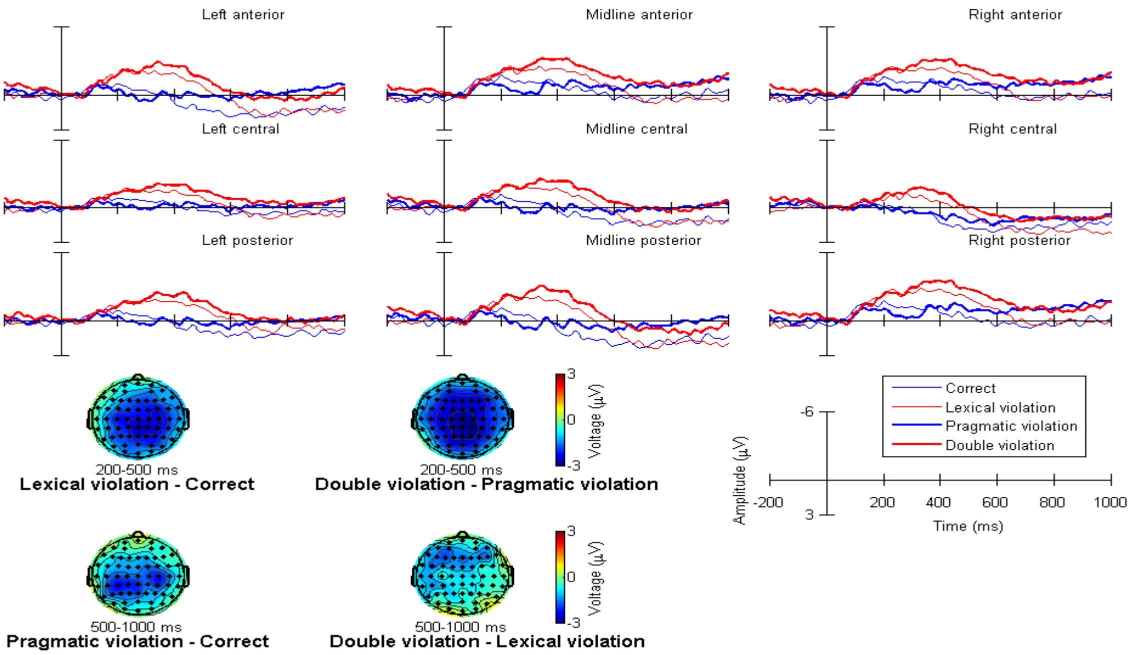


Figure 7. Effects of lexical and pragmatic inconsistency in Experiment 2. Upper portion: Grand average ERPs (a 30 Hz low-pass filter was applied for plotting) at nine scalp regions. Lower portion: Topographic maps of difference waves.

2.3.2.2.1. Quantifier position

We quantified the effects of pragmatic and semantic inconsistency using the mean ERP amplitudes over the 300-1000 ms window. There was a significant interaction of Quantifier and Consistency, reflecting the fact that inconsistent *some of* elicited a negativity (lateral: $F(1,19) = 8.03$, $p = .011$; midline: $F(1,17) = 3.59$, $p = .073$) whereas inconsistent *all of* elicited a positivity (lateral: $F(1,19) = 7.72$, $p = .012$; midline: $F(1,17) = 5.63$, $p = .028$). There was also a significant interaction of Quantifier, Consistency, and Region in the lateral ANOVA only. The interaction was due to the fact that the negativity for the *some of* sentences was broadly distributed (the Consistency by Region interaction for *some of* did not reach significance, $F(5,95) < 1$), whereas the positivity for the *all of* sentences was somewhat left-posterior in distribution. Specifically, for semantically inconsistent *all of* sentences, the Consistency by Region interaction was significant ($F(5,95) = 2.80$, $p = .033$); the simple effect of semantic Consistency was significant in the left posterior (*p* = .001), right posterior (*p* = .005), and left central (*p* = .046) regions, and marginal in the left anterior (*p* = .063) and right central (*p* = .054) regions.

2.3.2.2.2. Object position

N400. The N400 was quantified using mean amplitudes in the 200-500 ms time window. In this window there was a highly significant effect of Lexical Consistency, reflecting the fact that both lexically inconsistent conditions (picture-sentence mismatch and double violation) elicited more negative ERPs than lexically consistent conditions (correct object, and correct object following a pragmatically inconsistent quantifier). The effect was broadly distributed (it did not interact significantly with Region). The effect of Pragmatic Consistency was not significant. Crucially, no interactions of Pragmatic Consistency and Lexical Consistency were significant, indicating that the presence of a pragmatic violation did not modulate the lexico-semantic N400 effect.

Effect	200-500 ms	500-1000 ms
Pragmatic Consistency	*F*(1,19) = 0.36 *F*(1,19) = 0.45	*F*(1,19) = 22.96**** *F*(1,19) = 23.76****
Lexical Consistency	*F*(1,19) = 58.82**** *F*(1,19) = 53.15****	*F*(1,19) = 0.21 *F*(1,19) = 0.06
Region	*F*(5,95) = 60.48**** *F*(2,38) = 54.18****	*F*(5,95) = 29.46**** *F*(2,38) = 27.64****
Pragmatic Consistency × Lexical Consistency	*F*(1,19) = 0.60 *F*(1,19) = 0.60	*F*(1,18) = 0.27 *F*(1,19) = 0.19
Pragmatic Consistency × Region	*F*(5,95) = 0.57 *F*(2,38) = 1.88	*F*(5,95) = 1.24 *F*(2,38) = 0.48
Lexical Consistency × Region	*F*(5,90) = 1.38 *F*(2,38) = 2.59	*F*(5,95) = 1.05 *F*(2,38) = 0.59
Pragmatic Consistency × Lexical Consistency × Region	*F*(5,90) = 0.30 *F*(2,38) = 0.15	*F*(5,95) = 2.26* *F*(2,38) = 1.73

Table 3. Results of the lateral and midline omnibus ANOVAs at the object position in Experiment 2, with each cell showing the lateral ANOVA result above and the midline ANOVA result below. *.05 < p < .1; **p < .05; ***p < .005; ****p < .001.

Late negativity. The late ERP effect was quantified using the mean amplitudes in the 500-1000 ms window. In this window there was a significant main effect of Pragmatic Consistency, indicating that objects following pragmatic violations elicited more negative ERPs in the late window. In the lateral ANOVA there was a marginal interaction between Pragmatic Consistency, Lexical Consistency, and Region, due to the fact that although the main effect of pragmatic inconsistency was significant for both lexically correct (i.e., correct objects following pragmatically inconsistent quantifiers) and lexically incorrect (i.e., double violations) sentences, it was somewhat broadly distributed for lexically correct sentences (the interaction of Pragmatic Consistency and Region did not reach significance in these sentences, *F*(5,95) = 1.20, *p* = .320), but was more limited to the anterior regions for the double violations. Specifically, for the double violations, the interaction of Pragmatic Consistency and Region was marginally significant

(*F*(5,95) = 2.23, *p* = .095), and the Pragmatic Consistency effect was significant or marginal in the left anterior (*p* = .017), right anterior (*p* = .034), and left central region (*p* = .070), but not significant in the right central, left posterior, or right posterior regions (*p*s > .190).

To investigate whether the topographical difference was likely to be due to qualitatively different underlying sources or to quantitative differences in the signal, a scaling analysis was performed (Jing et al., 2006), which tests whether the signal in one effect has the same topography as the signal in another effect after being scaled based on a hypothetical scaling factor that represents the change in signal that would occur from a quantitative change in the strength of the underlying source. In this analysis, in which the pragmatic effect for the double violation (formed by subtracting the ERP for the mismatching object condition from the ERP for the double violation condition) was directly compared to that for the pragmatic violation (subtracting the correct condition from the pragmatic violation), the interactions with region were not significant (*F*(5,95) = 1.60, *p* = .204; *F*(5,95) = 1.85, *p* = .147),[14] indicating that the topographic differences found in the raw analysis are not likely to result from different underlying generators.

2.3.3. Discussion

At the quantifier position, the finding of Experiment 1 was partially replicated: pragmatic violations elicited a sustained negativity, albeit broader in distribution than the effect in Experiment 1. Unlike Experiment 1, semantically and pragmatically inconsistent quantifiers did not elicit similar effects in any time window; also unlike Experiment 1, a sustained positivity was

[14] In the procedure proposed by Jing and colleagues (2006), it is recommended to perform two comparisons: one between the raw Condition 1 and the scaled Condition 2, and one between the scaled Condition 1 and the raw Condition 2. Therefore, two *F*-tests are reported here.

observed for the semantically inconsistent quantifiers. The primary differences between the experiments were stimulus presentation modality (auditory in Experiment 2, visual in Experiment 1), task (consistency rating in Experiment 2, correctness judgments and comprehension questions in experiment 1), and composition of other sentences in the experiment (in particular, Experiment 1 did not include sentences with both pragmatic and lexico-semantic violations).[15] Importantly, in both experiments semantically inconsistent quantifiers elicited a qualitatively different ERP pattern than the pragmatically inconsistent quantifiers, which provides evidence that the sustained negativity for pragmatically inconsistent quantifiers does not reflect a general reanalysis mechanism or a general response to unexpected input, but rather a process specific to the kinds of revision or inhibition processes that are necessary for revising/inhibiting the pragmatic interpretation of a quantifier and activating its semantic meaning.

The ERPs elicited at the object position showed evidence that pragmatic and lexico-semantic information were processed independently: the presence or absence of a pragmatic violation upstream did not modulate the lexico-semantic N400 effect elicited by picture-sentence mismatch. The lack of an interaction cannot be explained by assuming that pragmatic revision had already been completed by the time the object was heard, since the objects still elicited a sustained negativity associated with pragmatic revision. Rather, the finding

[15] It should also be noted that, whereas Quantifier and Consistency were fully crossed in a Latin square design in Experiment 1, in Experiment 2 they were not: the items for the *some of* conditions (pragmatically inconsistent and correct *some of*) came from the 160 critical items, whereas the items for the *all of* conditions came from 80 fillers. Thus, while the Consistency factor was balanced across lists (i.e., a given item appeared in inconsistent conditions in some lists and consistent conditions in others), the Quantifier factor was not (it was not the case that, for a given item, it appeared with *some of* in some lists and *all of* in others). This was necessary for the experimental design; creating a fully crossed 2×2×2 design would have required more stimuli than it would have been feasible to create and to show to a single participant (assuming 40 trials per condition, it would have required 320 critical trials and at least as many fillers).

suggests that the revision or inhibition of the pragmatic interpretation of scalar terms utilizes different processing resources than those used for lexico-semantic prediction and integration. The late time window on the ERPs time-locked to objects continued to show a sustained negativity in response to pragmatically inconsistent sentences, suggesting that pragmatic revision was not yet completed by the time the object was encountered (which was, on average, 1300 ms after the onset of the quantifier). Thus, the data seem to suggest that pragmatic and semantic aspects of meaning were processed in parallel and their respective effects were additive.

2.4. EXPERIMENT THREE

The sustained negativity elicited by pragmatically inconsistent quantifiers in Experiment 1 was significant but small. While a similar but larger effect was observed in Experiment 2, the design of that experiment was different. Therefore, the main purpose for Experiment 3 was to attempt to replicate the negativity observed in Experiment 1 while using the same design (crossing Quantifier and Consistency) and only minor differences in the stimuli (see Materials, below). In addition, Experiment 3 tested whether group differences would emerge at the quantifier position, as they did at the quantified noun in recent ERP studies (Nieuwland et al., 2010; Hunt et al., 2013). In order to address this question, several measures of participants' judgments were collected in an offline questionnaire (Appendix A:), in addition to the consistency ratings collected during the recording.

2.4.1. Methods

2.4.1.1. Participants

Thirty-two Peking University students (15 females; mean age 22.6 years, range 18-28) who were native speakers of Mandarin participated in the study. Five were excluded from the statistical analysis due to excessive artifacts in their recordings, leaving a total of 27 participants in the final analysis. All participants had normal or corrected-to-normal vision and were right-handed according to the Chinese Handedness Survey (Li, 1983). All participants provided their informed consent and received payment, and all methods for the study were approved by the Ethics Committee of the Department of Psychology, Peking University, and the Human Subjects Committee of Lawrence at the University of Kansas.

2.4.1.2. Materials

The pictures used were identical to those used in the previous experiments. The sentences used in the current experiment followed the same criteria as those used in Experiment 1, with the exception that no extra phrase was included after the object in cases where having no extra phrase made the sentence sound more natural. As in Experiment 1, "some"- and "all"-type pictures were crossed with "some"- and "all"-type sentences to form a 2x2 experiment comparing the effects of pragmatic violations (relative to matched controls) to those of logic violations (relative to matched controls).

One hundred sixty picture-sentence pairs were used as fillers. The filler pictures met the same specifications as the critical trials, except that some of them depicted intransitive events. Eighty of the fillers were correct sentences (40 each of correct "some"-type and "all"-type picture-sentence pairs), 40 consisted of sentences with objects that did not match the objects shown in the picture (20 each of "some"-type and "all"-type pairs), and 40 consisted of sentences with verbs that did not match the activities shown in the picture (20 each of "some"-type and "all"-type pairs). Unlike in Experiment 1, all fillers used the same quantifiers as the critical sentences (*you-de* 有的 "some of", and *suoyou-de* 所有的 "all of"), to eliminate the possibility that subjects might recognize the critical quantifiers as a cue that there was no object or verb mismatch error coming up in the sentence.

All the experiment sentences were read aloud by a female Mandarin speaker from Beijing The recording was carried out within an anechoic chamber at the University of Kansas, using an ElectroVoice 767 microphone and a Marantz PMD-671 digital solid-state recorder sampling at 22050 Hz and in mono format.

2.4.1.3. Offline questionnaire

Participants completed a paper-pencil questionnaire in Chinese after the ERP recording. The full questionnaire is shown in Appendix A: (with an English translation in Appendix B:). For the purposes of data analysis, only responses to questions 2, 3, and 5 were used. Question 2 showed participants an underinformative picture-sentence combination (*All*-type picture with *some of* sentence) and asked them to qualitatively describe whether the sentence and picture were consistent, and why. Responses to this question were coded as "pragmatic" (participants who found the combination inconsistent or "not totally consistent") or "logical". Question 3 showed participants a *some of* sentence and five pictures (in which five, four, three, two, or one out of the five characters were performing the action described), and asked them to indicate all pictures that were consistent with the sentence. Responses were coded as "pragmatic" if the participant did not select the five-out-of-five picture, and "logical" if the participant did. Finally, Question 5 presented participants with a series of category sentences in the style of Noveck and Posada's (2003) stimuli, which were either true (e.g., *Some buildings have elevators*) or underinformative (e.g. *Some sentences have words*). Participants were asked to rate, on 1-7 scales, both the truth and the naturalness of the sentences. For the purpose of an individual difference measure, only their responses on the truth judgment of underinformative sentences were analyzed. Assuming that such sentences are true (logically speaking) but pragmatically infelicitous and thus unnatural, then participants who give high truth ratings for these sentences were assumed to be better at realizing the semantic meaning as separate from the pragmatic meaning, whereas those who give low truth ratings to these sentences were assumed to be poor at distinguishing between the semantic and pragmatic meanings.

2.4.1.4. Procedure

The procedure for the EEG recording was identical to that of Experiment 2, except for the following changes. 1) The experiment was divided into two sessions, with half of the stimuli being presented at each session; the first session began with a practice block consisting of 10 sentences, and the second session began with a practice block consisting of five sentences; the offline questionnaire was completed at the end of the second session. 2) The fixation point appearing at the beginning of the trial remained on-screen for 250 ms, and the picture remained on-screen for 5000 ms. 3) The delay between the offset of the picture and the onset of the sentence varied from 250-750 ms. 3) The rating scale appeared immediately at the offset of the sentence, rather than 100 ms later. 4) Each session of the experiment included 160 trials, divided into four blocks of 40 trials each; after every 20 sentences the participant was given a 10-second break, and after every 40 sentences the participant was given a full break. 5) The lists and fillers were pseudorandomized with the constraints that no more than four correct or incorrect sentences could appear consecutively, no more than four "some"- or "all"-type sentences could appear consecutively, and no more than four "some"- or "all"-type pictures could appear consecutively. 6) The recording took about 50 minutes.

2.4.1.5. Data acquisition and analysis

Data were acquired using the same equipment and settings as in Experiment 2, except that the sampling rate was 1000 Hz. Offline data analysis followed the same procedure as the data analysis for the quantifiers in Experiment 2, except that 1) a 0.5 Hz high-pass filter was applied to the continuous data before any other procedures were performed (this was to attenuate low-frequency skin potentials, which many trials were contaminated by); 2) ICA decompositions were performed separately for each session of each participant's recording, after which the two sessions for each participant were combined; 3) artifact rejection was only performed after, not before, ICA decomposition; 4) to allow for correlating with between-subjects covariates, ERP voltages were converted into z-scores using each participant's mean and standard deviation over all scalp channels.[16]

Data were analyzed statistically using repeated measures ANOVAs with the within-subjects factors Consistency (consistent, inconsistent), Quantifier (*some of*, *all of*), and Region (6 levels for the lateral ANOVA, 3 for the midline ANOVA).[17] For exploratory analysis, mixed ANOVAs using between-subject measures (average online consistency ratings, and the offline questionnaire measures described above) were also conducted, with the α level set to .01. The Huynh-Feldt correction was applied to *F*-tests with more than one degree of freedom in the numerator.

[16] Analyses were also conducted using raw ERP voltages rather than z-scores. Where the two analysis methods yielded different results, the discrepancies are noted in the text.

[17] Exploratory analyses of the raw voltages were also conducted using linear mixed models with crossed random factors for subjects and items (Baayen et al., 2008). Model significance was evaluated using log-likelihood tests and the significance of coefficients using Markov chain Monte Carlo sampling. The α level was set to .01. Where the two analysis methods yielded different results, the discrepancies are noted in the text.

2.4.2. Results

2.4.2.1. Behavioral results

The participants' task was to rate the consistency between the picture and the sentence using a 7-point scale. One participant's data were not saved because of a software error, and thus the analysis was conducted based on data from 26 participants. Average ratings are shown in Figure 8. A repeated measures ANOVA revealed significant effects of Consistency ($F(1,25) = 250$, $p < .001$) and of Quantifier ($F(1,25) = 107.2$, $p < .001$), and a significant interaction ($F(1,25) = 167.2$, $p < .001$). Planned t-tests revealed that for both quantifier types, inconsistent sentences received lower ratings than consistent sentences (*some of*: $t(25) = -4.51$, $p < .001$; *all of*: $t(25) = -24.73$, $p < .001$), and that the interaction was due to the effect of inconsistency being more pronounced for *all of* sentences than for *some of* sentences ($t(25) = 12.93$, $p < .001$). Nine out of 26 participants reliably related pragmatically inconsistent sentences lower than consistent *some of* sentences (ps < .05), and 25 out of 26 reliably rated these sentences higher than semantically

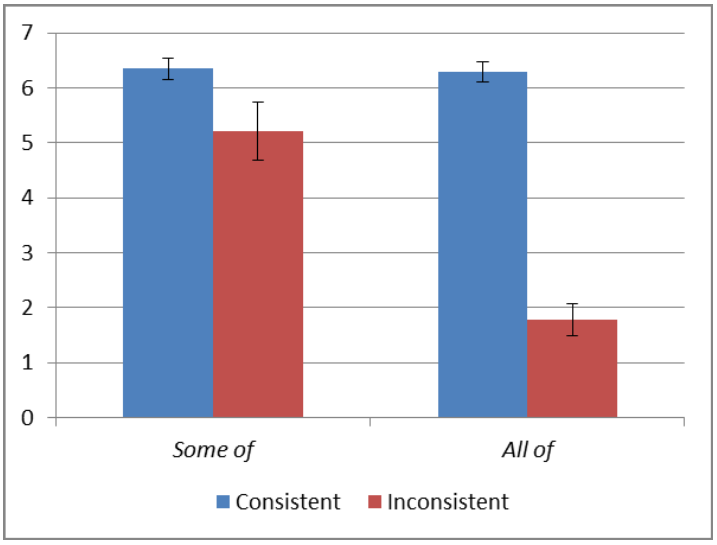


Figure 8. Consistency ratings in Experiment 3 (1=very inconsistent, 7=very consistent). Error bars represent ±2 standard errors of the mean.

inconsistent sentences (*p*s < .05).

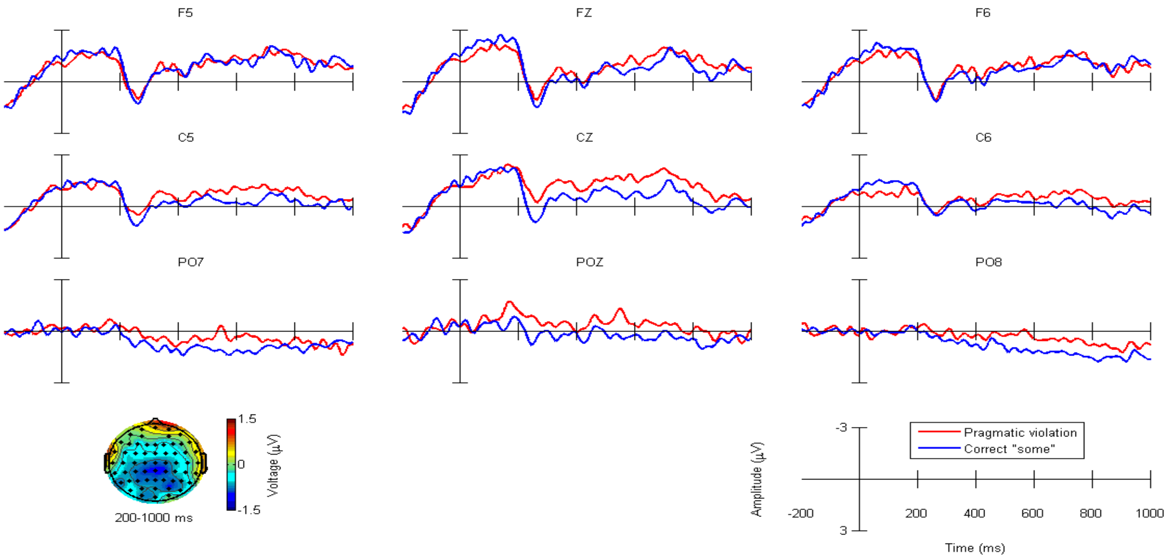


Figure 9. Effect of pragmatic inconsistency in Experiment 3. Upper portion: Grand average ERPs (a 30 Hz low-pass filter was applied for plotting) at nine scalp regions. Lower portion: Topographic map formed by subtracting the correct *some of* condition from the pragmatically incorrect condition.

2.4.2.2. ERP results

ERPs elicited by *some of* are depicted in Figure 9, and those elicited by *all of* in Figure 10. Visual inspection of the data suggests that pragmatically inconsistent *some of* elicited a sustained negativity, broadly distributed over centro-parietal sites, which began approximately 200 ms after the onset of the quantifier and lasted through the rest of the epoch. Semantically inconsistent *all of*, on the other hand, did not elicit a clear pattern, in this time window, although it appeared to elicit an increased negativity over right anterior sites from 300 to 500 ms. Therefore, two time windows were analyzed: one from 300-500 ms, and one from 500-900 ms. Below the results of the whole-group analysis without between-participant covariates are reported first, followed by the results of the analysis using the between-participant covariates.

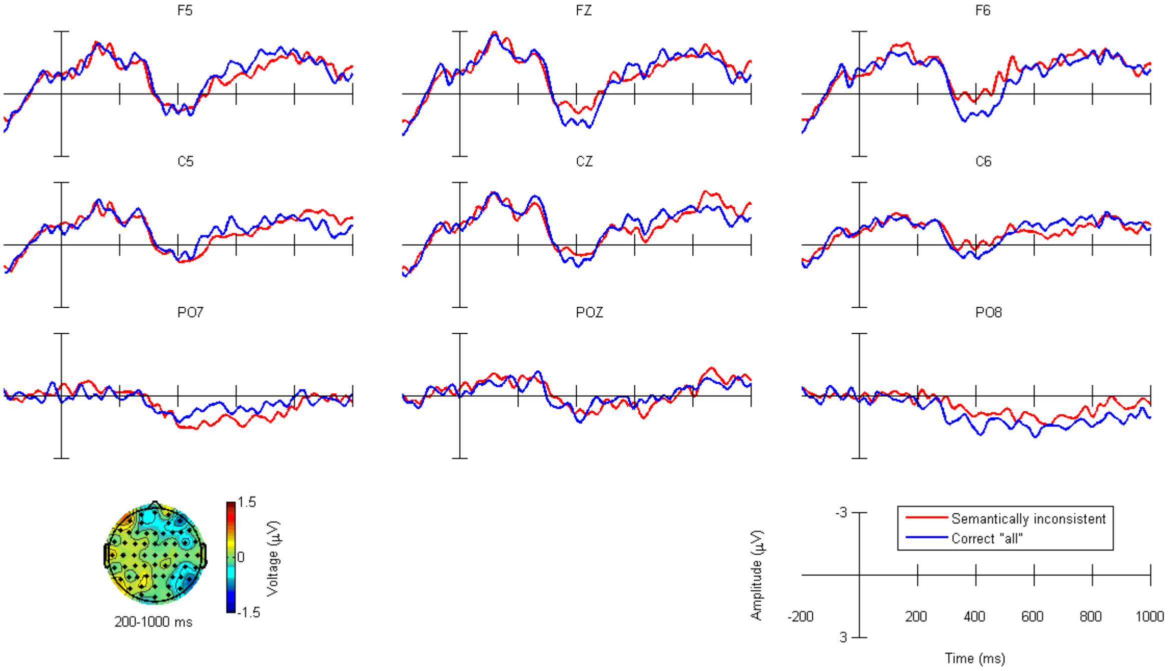


Figure 10. Effect of semantic inconsistency in Experiment 3. Upper portion: Grand average ERPs (a 30 Hz low-pass filter was applied for plotting) at nine scalp regions. Lower portion: Topographic map formed by subtracting the correct *all of* condition from the semantically incorrect condition.

2.4.2.2.1. Overall analysis

300-500 ms. The ANOVA on the z-scores averaged over this time window revealed marginal main effects of Consistency (lateral: $F(1,26) = 3.90$, $p = .059$; midline: $F(1,26) = 3.58$, $p = .070$), indicating that both pragmatically inconsistent and semantically inconsistent quantifiers elicited negativities in this time window.[18] Exploratory linear mixed effects models yielded a significant three-way interaction of Consistency, Quantifier, and Region ($\chi^2(8) = 29.65$, $p < .001$). This analysis suggested that pragmatically inconsistent *some* yielded significant negativities (*p*s < .001) in the left central, left posterior, midline central, and right posterior regions, and marginal negativities (*p*s < .05) in the left anterior and midline posterior regions;

[18] When the ANOVA was conducted on raw voltages rather than z-scores, no significant effect of Consistency or interactions with Consistency were obtained.

whereas semantically inconsistent *all* yielded a significant negativity in the right anterior region only ($p < .001$), marginal negativities in the right central and posterior regions (ps < .02), and a nearly marginal negativity in the midline anterior region ($p = .062$).

In sum, both types of inconsistency yielded negativities in this time window, although exploratory analyses using mixed-effects modeling suggest that the negativity elicited by pragmatic inconsistencies had a broad posterior distribution whereas the negativity elicited by semantic inconsistencies had a right frontal distribution.

500-900 ms. The ANOVA on the z-scores averaged over this time window revealed a marginal interaction between Quantifier and Consistency in the lateral analysis (lateral: $F(1,26) = 3.27, p = .082$; midline: $F(1,26) = 2.75, p = .109$).[19] This interaction was due to an effect of Consistency in the *some of* sentences ($F(1,26) = 5.87, p = .023$) but not the *all of* sentences ($F < 1$). The same pattern of results was observed, with a higher significance level, in the exploratory linear mixed effects analysis: the Quantifier*Consistency interaction was significant ($\chi^2(1) = 103.2, p < .001$), and the effect of inconsistency was non-significant for *all of* sentences ($t = 0.21$, $p > .8$) but significant for *some of* sentences ($t = -10.16, p < .001$). The three-way interaction of Quantifier, Consistency, and Region was not significant ($\chi^2(8) = 1.83, p = .159$).

In sum, during this time window a negativity was only observed for pragmatically inconsistent *some of*, and it had a broad distribution.

2.4.2.2.2. Analysis with individual-level covariates

[19] When the ANOVA was conducted on raw voltages rather than z-scores, the interaction reached significance in the midline analysis ($F(1,26) = 4.44, p = .045$) but did not approach significance in the lateral analysis ($F(1,26) = 2.53, p = .124$).

The goal of this analysis was to identify participant-level predictors that showed at least third-order interactions with Quantifier and Consistency (as this is the minimum interaction necessary to establish that a predictor specifically affects responses to pragmatic violations). Because of the exploratory nature of the analysis, the α level was set to .01. The three individual-level predictors were whether the participant responded pragmatically on Question 3 (see section 2.4.1.3, "Offline questionnaire"), hereafter referred to as Group; the average truth value the participant gave to underinformative sentences in the offline questionnaire, hereafter referred to as Truth Rating; and the difference between consistency ratings the participant gave to pragmatically inconsistent and consistent *some of* sentences during the recording, hereafter referred to as Consistency Rating. Some participants did not respond the same way to both questions 2 and 3, as shown by the cross-tabulation in Table 4. Nevertheless, a logistic regression showed that across participants, responses to one question were significantly predictive of responses to the other (*b* = 2.28, SE = 0.94, *z* = 2.43, *p* = .015). Therefore, to minimize the number of comparisons, and because two participants did not complete question 3, only responses to question 2 were used in the analysis reported below; the same pattern of results was also found when using responses to question 3 instead.

		Response on Question 3	
		Pragmatic	**Logical**
Response on Question 2	**Pragmatic**	11	3
	Logical	3	8

Table 4. Cross-tabulation of responses to questions 2 and 3 on the offline survey. The cells sum to 25 rather than 27 because two participants failed to respond to Question 3.

Consistency Rating had a negative skew (-1.443), with most participants rating inconsistent *some of* sentences as slightly worse than consistent *some of* sentences, and only a few participants rating them as very much worse. Thus, before using this variable as a covariate, the values were reflected, transformed via reflected reciprocal, and then re-reflected, to transform the values into a more normal distribution (skew = -.006). Since one participant's consistency ratings were not saved, that participant was excluded from the analyses. Truth Rating had a roughly bimodal distribution with peaks at the endpoints (1 and 7), as shown in Figure 11. Both of these variables were centered before statistical tests were conducted.

No predictors showed interactions of interest in the ANOVAs on either time window.

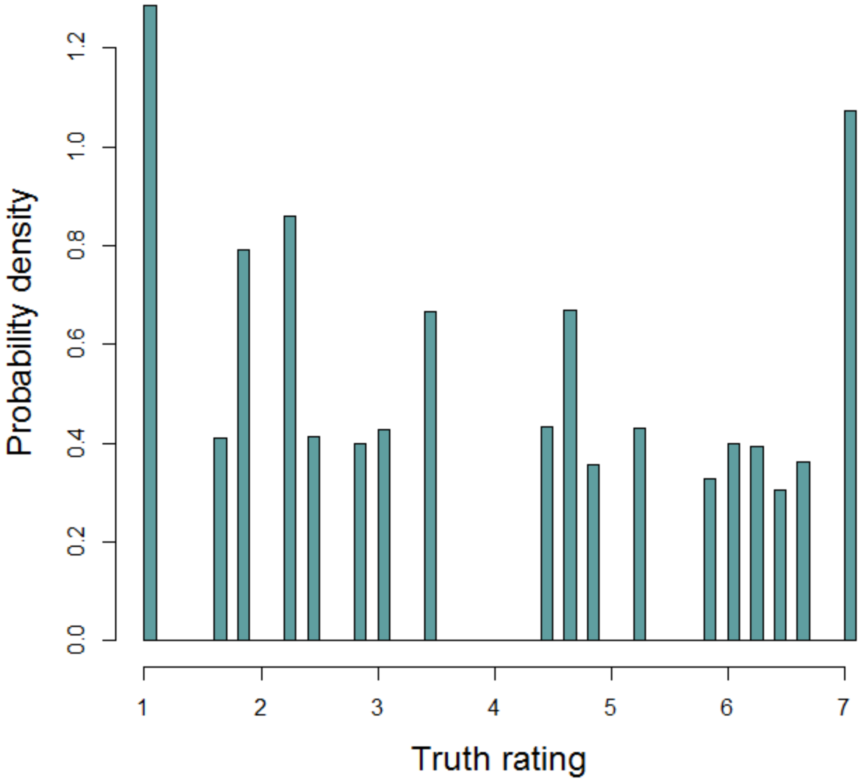


Figure 11. Distribution of truth ratings of underinformative sentences in the offline survey.

Exploratory analyses using linear mixed effects models, however, showed several interactions. These are discussed below.

300-500 ms. The linear mixed effects model revealed a significant three-way interaction between Consistency, Quantifier, and Consistency Rating ($\chi^2(1) = 10.78$, $p = .001$), as well as a significant four-way interaction between Consistency, Quantifier, Region, and Group ($\chi^2(8) = 21.13$, $p = .006$).

The three-way interaction with Consistency Rating is illustrated in Figure 12. Resolving the interaction by Quantifier showed that there were significant interactions of Consistency with

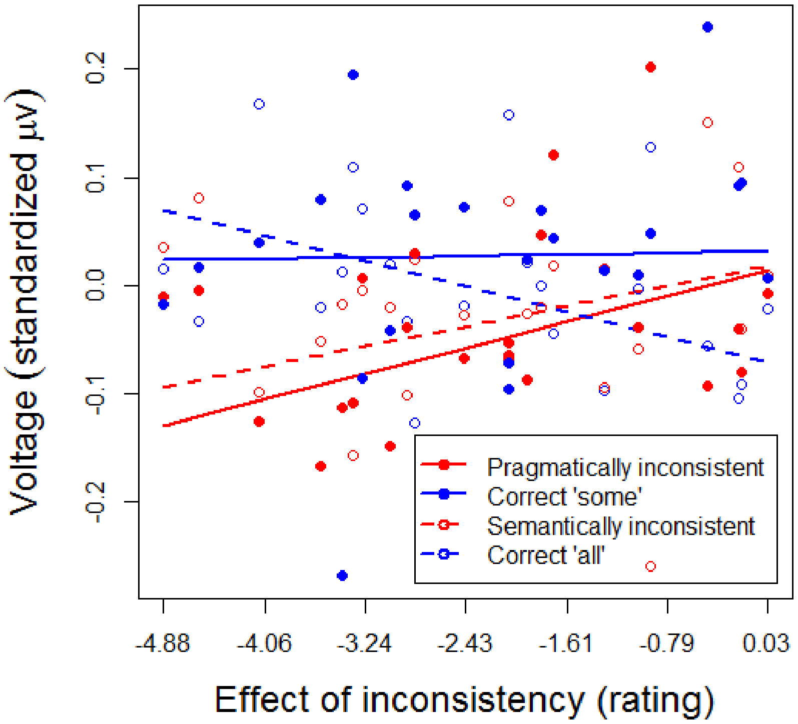


Figure 12. Relationship between Quantifier, Consistency, and participants' sensitivity to pragmatic inconsistency. The x-axis shows the participant's sensitivity in the online ratings (consistency ratings for pragmatically inconsistent *some of* sentences minus ratings for correct *some of* sentences; more negative values indicate greater sensitivity). The y-axis shows ERP voltages.

Consistency Rating for both *some of* sentences ($\chi^2(1) = 40.97$, $p < .001$) and *all of* sentences ($\chi^2(1) = 106.38$, $p < .001$). As shown in Figure 12, for both types of sentences it was the case that the negativity elicited by inconsistent quantifiers was largest for participants who rated pragmatically inconsistent sentences as much worse than correct *some of* sentences, whereas participants who rated these sentences as similar tended to show less negativity, or even a positivity, for both pragmatically inconsistent quantifiers and semantically inconsistent quantifiers. The main difference between the pragmatic and semantic violations was the magnitude of the negativity—pragmatically inconsistent sentences tended to show negativities for all subjects, whereas semantically inconsistent sentences tended to show positivities for participants who showed the least sensitivity to the pragmatic inconsistency in their ratings.

As for the interaction with Group, resolving the interaction by Quantifier revealed that *all of* sentences showed no significant interactions with Group (*ps* > .4), whereas *some of* sentences showed a significant three-way interaction between Consistency, Region, and Group ($\chi^2(1)$ =

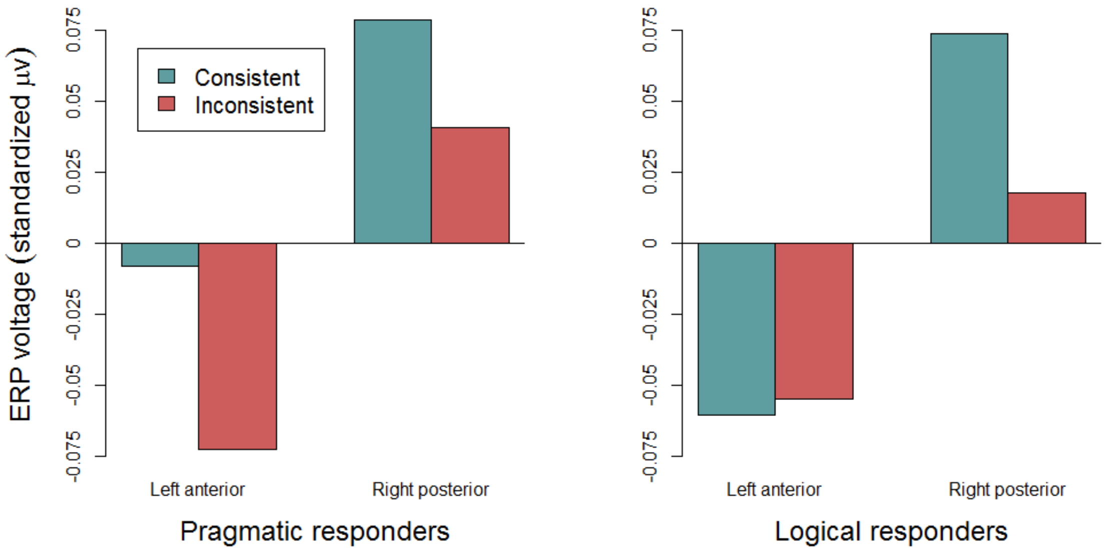


Figure 13. Relationship between Consistency, Group (pragmatic responders vs. logical responders), and ERP responses to *some of* at two scalp regions. The y-axis shows the mean ERP voltage over 300-500 ms, in z-score standardized µv.

40.97, $p < .001$). Further resolving this interaction by Region revealed significant interactions between Consistency and Group in the left anterior region ($\chi^2(1) = 17.7$, $p < .001$) and right posterior region ($\chi^2(1) = 10.11$, $p = .001$). These interactions are shown in Figure 13. As the figure indicates, pragmatic responders (those who responded "Inconsistent" on Question 2 of the offline survey) showed a negativity with a different topography than logical responders: for pragmatic responders the negativity extended into the left anterior region, whereas for logical responders it did not. (All other regions except for midline anterior showed significant effects of Consistency (*p*s < .009), suggesting that both groups showed negativities in those regions.)

500-900 ms. The linear mixed effects model revealed significant three-way interactions between Consistency, Quantifier, and Truth Rating ($\chi^2(1) = 12,71$, $p < .001$), and between Consistency, Quantifier, and Group ($\chi^2(1) = 11.29$, $p < .001$). The four-way interaction between Consistency, Quantifier, Region, and Group was also significant ($\chi^2(8) = 24.71$, $p = .002$).

The three-way interaction of Consistency, Quantifier, and Truth Rating is plotted in Figure 14. Resolving this interaction by Quantifier revealed significant interactions of Consistency and Truth Rating for both *some of* ($\chi^2(1) = 104.22$, $p < .001$) and *all of* sentences ($\chi^2(1) = 23.65$, $p < .001$). As shown in the figure, for both *some of* and *all of* sentences, the

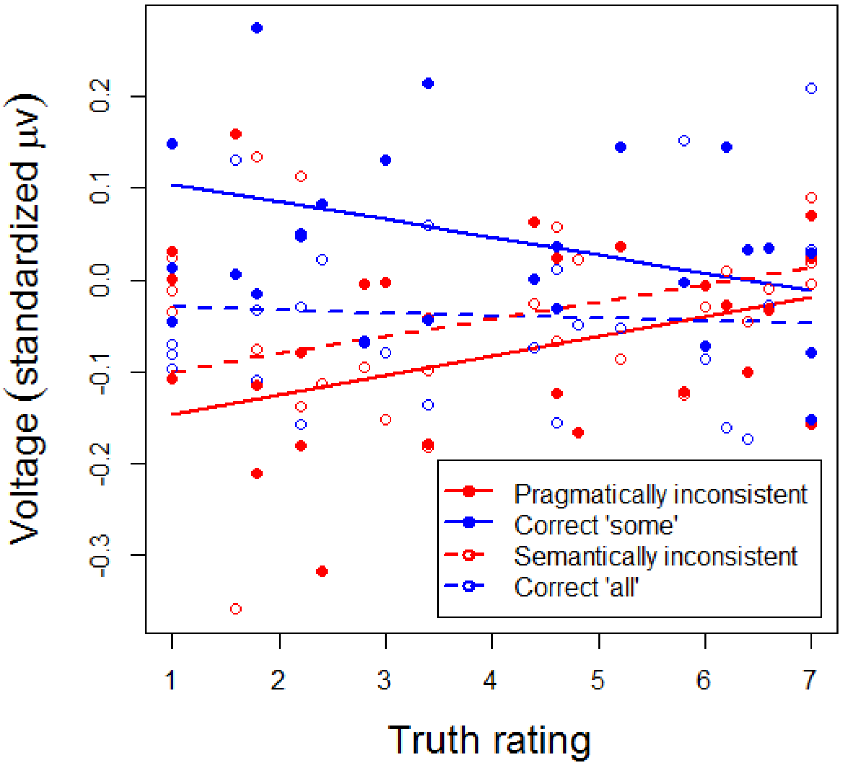


Figure 14. Relationship between Quantifier, Consistency, and participants' offline truth ratings of underinformative sentences. The x-axis shows the participant's truth ratings (higher values suggest better ability to realize semantic interpretations). The y-axis shows ERP voltages.

negativity elicited by inconsistent quantifiers tended to be largest for participants who were poor at distinguishing between the semantic and pragmatic meanings (those who judged underinformative sentences as "untrue"). This pattern appeared for both *some of* and *all of* sentences, but with different magnitudes; as is apparent from the figure, to effect of inconsistency in *some of* sentences ranged from negative (among participants who gave low truth ratings) to no effect (among participants who gave high truth ratings), whereas the effect of inconsistency in *all of* sentences was a positivity among some participants who gave high truth ratings.

Resolving the four-way interaction of Consistency, Quantifier, Region, and Group revealed that there was a significant Consistency × Group × Region interaction for *some of* sentences ($\chi^2(8) = 27.26$, $p < .001$) but not for *all of* sentences ($\chi^2(8) = 8.07$, $p =.427$). In *some of* sentences there were significant Consistency × Group effects in several regions (left anterior: $\chi^2(1) = 4.15$, $p = .042$; right central: $\chi^2(1) = 4.04$, $p = .044$; left posterior: $\chi^2(1) = 5.95$, $p = .015$; right posterior: $\chi^2(1) = 15.67$, $p < .001$), indicating that in each of these regions one group showed a greater sustained negativity than the other. The interaction is plotted in Figure 15. As the figure indicates, over posterior sites the negativity was greatest for logical responders, whereas over left anterior sites it was greatest for pragmatic responders; this suggests that the two different groups showed effects with differing scalp distributions.

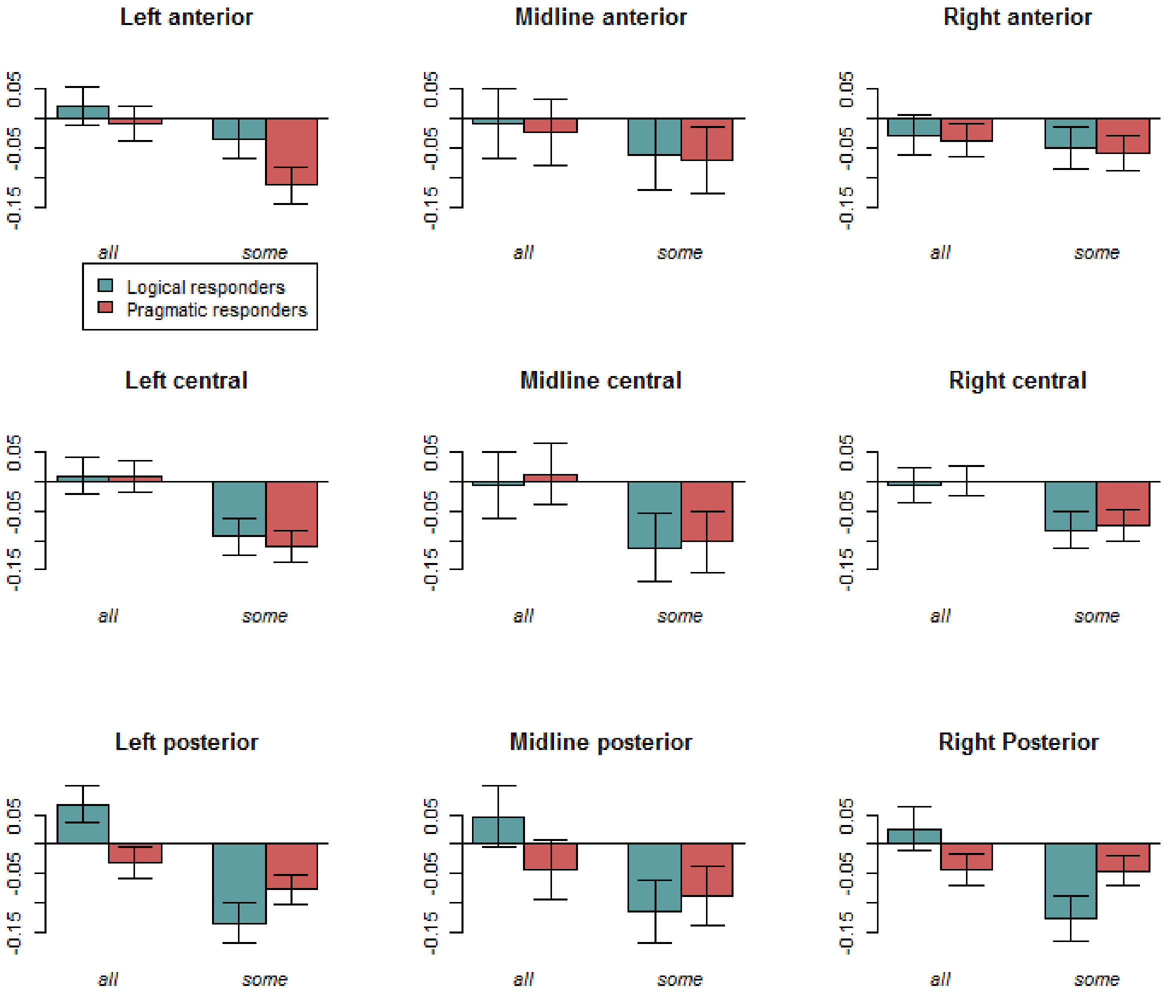


Figure 15. Relationship between Consistency, Quantifier, Group (pragmatic responders vs. logical responders), and ERP responses. The y-axis shows the ERP effect (mean voltage over 500-900 ms, in z-score standardized µv, for pragmatically inconsistent *some of* or semantically inconsistent *all of*, minus mean voltage over 500-900 ms for correct *some of* or correct *all of*). Blue bars show effects for logical responders, and red bars effects for pragmatic responders.

2.4.3. Discussion

The purpose of Experiment 3 was to replicate the comparison between pragmatic and semantic inconsistencies from Experiment 1, and to examine potential individual differences in the sustained negativity elicited by quantifiers that are pragmatically infelicitous with a context.

Regarding the first goal, such quantifiers in this experiment once again elicited a broad sustained negativity, replicating the findings of the previous experiments. The ERP elicited by semantically inconsistent quantifiers, on the other hand, differed from that elicited in the previous experiments (in Experiment 1 there was only an early anterior positivity, in Experiment 2 there was a sustained broad positivity, and in this experiment there was an early, right-anterior negativity). Nevertheless, this effect was dissociated from the ERP elicited by pragmatically inconsistent quantifiers, both in terms of morphology (pragmatically inconsistent quantifiers elicited a negativity with a more sustained duration) and topography (pragmatically inconsistent quantifiers elicited a negativity with a broad posterior distribution, whereas semantically inconsistent quantifiers elicited one with a more restricted right-anterior distribution).

Regarding the second goal, the relationship between the individual-level predictors tested and the sustained negativity remains inconclusive, but exploratory analyses provided preliminary evidence suggesting that 1) participants who tended to rate underinformative sentences as "false", or as inconsistent with a preceding picture, also tended to show greater negativities in response to underinformative quantifiers; and 2) the topography of the sustained negativity differed between participants who judged the pragmatically infelicitous sentences as inconsistent with their pictures in the offline and participants who judged these sentences as consistent. Regarding the first point, it is worth noting that participants who rate underinformative sentences as "true" and those who judge the pragmatically infelicitous pictures as "consistent" are both, presumably, participants who are either better at realizing the semantic meaning of the quantifier or are less sensitive to the pragmatic meaning (or both). These are precisely the participants who showed smaller negativities in response to the quantifier. This suggests that the sustained negativity may reflect extra effort that the participants who are poorer at realizing semantic meaning must spend to process the quantifier. The possible functional significance of the sustained negativity will be discussed in more detail in Section 2.5, General discussion.

The exploratory analyses also suggested that participants who tended to judge underinformative sentences as consistent with their contexts (as determined by the post-experiment questionnaire) showed ERP responses that differed in terms of topography from those of participants who judged underinformative sentences as inconsistent with their context. The latter group (pragmatic responders) showed a more left-anterior effect throughout the epoch, and the former group (logical responders) showed a more posterior effect than the pragmatic responders in the late portion of the epoch. The functional interpretation of this topographical difference was not predicted and is thus difficult to interpret, but it should be noted that an anterior negativity was observed in the experiment by Baggio and colleagues (2008), in which participants had to re-compute discourse models, but not in the experiment by Pijnacker and colleagues (2011), in which participants had to revise discursive inferences. This topographical difference might suggest functional differences in the way the two groups of participants in the present experiment—although it is not possible at this point to rule out the possibility that topographical differences between experiments could be due to different preprocessing routines used (in particular, Baggio and colleagues employed no ocular correction algorithm, Pijnacker and colleagues employed a regression-based algorithm, and in the present study ICA-based correction was used). It should also be noted that the authors of those studies do not propose that the difference in the topographies of the two effects necessarily imply functional differences between re-computing discourse models and revising discursive inferences.

2.5. GENERAL DISCUSSION OF ERP EXPERIMENTS

The three experiments reported here examined the neural responses to pragmatic violations while controlling for lexico-semantic factors and allowing for the detection effects at the moment the critical quantifier is encountered. Perhaps most importantly, different ERP patterns were observed for pragmatic and semantic violations: whereas lexico-semantic violations elicited an N400 and quantificational semantic violations elicited different effects in each experiment, pragmatic violations consistently elicited sustained negative components. Pragmatic violations were also the only conditions to elicit sustained negativities in these experiments. The results suggest that 1) the pragmatic reading of the quantifier is used rapidly during online processing and must be inhibited effortfully if it is inconsistent with the context; and 2) pragmatic inconsistency is processed differently than semantic inconsistency, at least in the context tested here. Experiment 2 also examined the interaction between pragmatic and lexico-semantic processing and found that pragmatic reanalysis did not modulate lexico-semantic processing downstream, suggesting that pragmatic and lexico-semantic aspects of meaning were processed independently. Below, each of these findings is discussed in turn.

2.5.1. The sustained negativity

At the quantifier position, in all experiments a sustained negativity was elicited by quantifiers that are pragmatically inconsistent with a context. This effect seems to be related to pragmatic processing in particular, as it was not elicited by quantifiers that were semantically inconsistent with a context. The effect could not be due only to processes related to seeing or hearing an unexpected word, since semantically inconsistent quantifiers and lexico-semantically inconsistent objects elicited qualitatively different effects even though they were also unexpected. The effect could also could not be due to revising expectations about what aspect of the picture will be pointed out later in the sentence, since this sort of revision is also possible in the semantically inconsistent *all of* sentences but did not elicit a sustained negativity. It is not likely to be due to generating or retrieving the pragmatic interpretation of the quantifier, since that process may have already been initiated during verbal pre-coding when the participant viewed each picture (Huang et al., 2010; Hartshorne & Snedeker, submitted). Rather, the sustained negativity is more likely to be related to effortful pragmatic reanalysis: specifically, inhibiting the pragmatic reading of *some of* and retrieving the semantic reading. This interpretation is consistent with several recent studies (Baggio et al., 2008; Pijnacker et al., 2011) that have observed sustained negativities related to revising discourse models or discourse-based inferences. Further support for this interpretation comes from a study by Leuthold and colleagues (2012), who observed a sustained right-posterior negativity (and corresponding left-frontal positivity) in response to emotion words that were incongruent with a situation previously described (e.g., "The golf pro was <u>distraught</u>", after a context suggesting that the golf pro had a good chance to win a tournament). They speculated that this negativity may be due to suppressing the expected emotion words. It is possible that such an operation also involves reconsideration of the character's point of view, which is a hallmark of Gricean pragmatic processing. While the linguistic manipulation in the present study is different than those discussed above, pragmatic violations in the present study would have led participants to reanalyze the implicature-based meaning of *some*, similar to Pijnacker et al. (2011), and to re-consider the point of view of another speaker or character, as in Leuthold et al. (2012). It is also possible that re-interpreting the quantifier requires constructing new mental models of the possible meanings of the quantified phrase in order to find a mental model that is consistent with the picture—i.e., a model in which some, and in fact all, of the girls are sitting on blankets. (For a review of mental models theory, see Nicolle, 2003).

It should be noted that an alternate strategy participants could employ to interpret sentences with inconsistent quantifiers is to make no attempt to evaluate the meaning and reference of the quantifier whatsoever until more information becomes available later in the sentence. Recall that semantic violations consisted of *Some*-type pictures (e.g., several girls sitting on blankets and the rest sitting on chairs) followed by sentences beginning "In the picture, all…". Such a sentence could turn out to be correct (e.g., "…all the girls are wearing bathing suits"), and thus it is possible that participants waited until they had more information before attempting to further evaluate the consistency between the sentence and the picture. Crucially, however, pragmatically inconsistent quantifiers could also be followed by sentences that turn out to be correct (e.g., an *All*-type picture could be followed by "…some of the girls are happy"). If participants employed such a processing strategy, one might expect effects to appear at the verb or object position, where the semantically incorrect sentences become unambiguously incorrect (e.g., at "…all the girls are sitting on…" or "…some of the girls are sitting on…", it becomes impossible to analyze the sentence as "…all the girls are wearing bathing suits" or "…some of the girls are happy"). Because the structure of the verbs used in the present study varied (verbs were presented simultaneously with aspect markers that preceded or followed them and differed in length and other properties) as did the point where the violation becomes unambiguous, such an analysis was not feasible with the present data, although the sustained negativity elicited by objects following pragmatic inconsistencies in Experiment 2 may be evidence for this sort of processing. Nevertheless, participants showed different ERP responses to the two types of inconsistency, even though this delayed interpretation strategy is available for both. Only the pragmatically inconsistent quantifiers can be reconciled with the context by reanalyzing the meaning of the quantifier (cancelling the implicature and retrieving the semantic meaning), and accordingly only the pragmatically inconsistent quantifiers showed the sustained negativity.

An alternative account of the sustained negativity observed in the present study is that it reflects truth-verificational processes initiated by the inconsistency between the quantifier and the context. Wiswede and colleagues (in press) found a late negativity elicited by nouns that make sentences untrue (e.g., "Africa is a planet"), but this negativity only occurred for participants who were performing a truth-value judgment task, not those who were performing a memory task. One might argue that pragmatically inconsistent *some of* in the present study initiated this truth-verificational process, whereas semantically inconsistent *all of* did not since its interpretation could be delayed until later in the sentence. Other aspects of the results, however, speak against this interpretation. In particular, no late negativity was elicited by objects that mismatched only the lexico-semantic content of the picture (e.g., the pure picture-sentence mismatch condition in Experiment 2, which only elicited an N400, as did the lexico-semantically mismatched objects in the Experiment 1 fillers in an exploratory analysis). Such words also introduce falsehood into the sentence, and are more similar to the words that elicited the late negativity in Wiswede and colleagues' (in press) study. Nevertheless, the sustained negativity in the present study only occurred in conditions where the inconsistency was related to pragmatic meaning.

The fact that the responses to the pragmatic condition were characterized by early recognition of the inconsistency and revision of the inference has implications for both the theory of scalar implicature processing and for the cognitive neuroscience of language; these implications are discussed below.

2.5.2. The costs of scalar implicature processing

The present set of experiments was not designed to test the time course and processing costs of realizing a pragmatic meaning (the processing cost question will be addressed in the next chapter), but it does provide evidence about the time course and costs of adjudicating between the semantic and pragmatic readings. As noted above, the sustained negativity effect at the quantifier position for conditions in which the pragmatic reading of the quantifier was inconsistent with the context suggests that suppressing that aspect of meaning and accessing the semantic aspect was costly and effortful. The fact that this effect is strongest in participants who are poor at distinguishing between the semantic and pragmatic interpretations (Experiment 3; see Figure 14) is consistent with this interpretation: retrieving the semantic reading may require more effort for these participants, making the sustained negativity more prominent. In line with this account, Feeney and colleagues (2004), based on findings from a speeded verification task, also concluded that participants reading underinformative instances of *some* needed to suppress the pragmatic meaning and that this suppression is cognitively taxing. Garrett & Harnish (2007) provide evidence from another pragmatic phenomenon, *standardization implicitures* (e.g., "I've had breakfast" is interpreted as "I've had breakfast today"), that the pragmatically enriched reading is computed by default and the semantic reading can only be retrieved with effort—although it is not necessarily the case that standardization-based implicitures are processed via the same mechanisms as scalar implicatures (see also Bezuidenhout & Cutting, 2002). On the other hand, a recent study in Mandarin suggests that the retrieval of the literal meanings of conventionalized lexical metaphors are not delayed relative to their metaphorical meanings (Lu & Zhang, 2012), raising the interesting possibility that pragmatic inferencing (at least scalar inference triggered by quantifiers) unfolds in a different manner than metaphor comprehension.

In sum, the results of the present study suggest that accessing the semantic reading of a scalar quantifier takes extra cognitive effort, eliciting a sustained negativity in the ERP. This is easy to reconcile with default models of scalar implicature processing (Levinson, 2000), which assume that inferences are realized quickly and with little regard for whether the enriched pragmatic meaning makes the sentence more informative, and that these inferences can be subsequently cancelled. It does not, however, preclude context-driven (Noveck & Sperber, 2007) or constraint-based models (Degen & Tanenhaus, 2011), since the possibility of verbal pre-encoding of the stimuli should have made the pragmatic reading easy to generate rapidly, and these models do not necessarily predict inhibition of pragmatic meaning to be effortless. Further study of the processing costs associated with both scalar inference realization and scalar inference reanalysis is needed to elucidate which cognitive resources are used for pragmatic processing and allow these models to become more explicit about this issue.

2.5.3. Neural correlates of different aspects of meaning processing

Much work on the processing of meaning in the brain has focused on the N400 ERP component and its sensitivity to manipulations of real-world plausibility (e.g., sentences such as "She spread her bread with socks"). Substantially fewer studies have examined how the brain processes compositional aspects of meaning (for reviews see Pylkkänen et al., 2011; Panizza, 2012) and how context and discourse interact with meaning (see Van Berkum, 2009). Scalar implicatures offer a promising test case for these issues, given that they represent an aspect of meaning that is composed in concert with semantic meaning and that the generation of scalar implicatures is strongly affected by context and expectations about speakers.

The present study offers converging evidence with other emerging work in neurosemantics suggesting that the mechanisms by which the brain composes meaning may not be the same as those by which it accesses words from the lexicon, notices associations between words, or evaluates real-world plausibility (i.e., several of the processes reflected by the N400). Recent investigations suggest that the patterns of brain activation elicited by violations of real-world plausibility are not the same as those elicited by linguistically-motivated abstract operations such as semantic composition (Pylkkänen et al., 2011), licensing of negative polarity items (Steinhauer et al., 2010; Panizza, 2012) and semantic subcategorization (Kuperberg et al., 2000). In the present experiments it was found that quantifiers which were pragmatically inconsistent with a context elicited a qualitatively different ERP response than quantifiers which were semantically inconsistent, suggesting that they were processed by different mechanisms. It was also found that costly pragmatic reanalysis of a quantifier's meaning did not modulate concurrent processing of lexico-semantic errors, providing further evidence that these two aspects of meaning are processed independently. It should be noted, however, that while the qualitative differences in ERP responses found in the present study are consistent with distinct mechanisms of pragmatic and semantic meaning composition, it is difficult to infer the underlying sources of the ERP pattern. For this reason, localizing the neural generators of these effects using methods with high spatial resolution would be a valuable avenue for further research, and could provide additional evidence for a dissociation of pragmatic and combinatorial semantic meaning composition.

2.5.4. Limitations and open questions

2.5.4.1. The baseline for comparison

In the present study, ERP responses to a quantifier that made a sentence pragmatically inconsistent with its context were compared against responses to a quantifier that made a sentence semantically inconsistent with its context. The goal of this manipulation was to isolate correlates of the processing of pragmatic inconsistency, while subtracting out other factors (such as the mismatch between the expected lexical item and the perceived one). Nonetheless, these two types of inconsistencies also differ in ways other than the presence or absence of a pragmatic interpretation. Particularly, even though *all of* and *some of* are often considered to belong to the same class of quantifiers (*logical quantifiers*; see, e.g., Morgan et al., 2011), the processes involved in verifying the meanings of *at least one of*, *not all of*, an *all of* may be different (see, e.g., Bott et al., 2012). To verify whether "*at least one of* the girls is sitting on blankets" (the semantic interpretation of "*some of* the girls are sitting on blankets") is true, the participant only needs to find one instance of set intersection (i.e., one entity in the context that is a girl and is sitting on a blanket). Verifying whether "*not all of* the girls are sitting on blankets" (the pragmatic interpretation of "*some of* the girls are sitting on blankets") is true (or failing to verify whether "*all of* the girls are sitting on blankets is true) requires a similar procedure, except that in this case the participant only needs to find one entity in the context that is a girl and is *not* sitting on a blanket. In either of these cases, once the participant finds one entity that meets the necessary criteria, she can in theory verify the meaning and stop examining entities (although it is an empirical question whether comprehenders actually do this in natural language). On the other hand, to verify whether "*all of* the girls are sitting on blankets" is true (or fail to verify "*not all of* the girls are sitting on blankets"), the participant must check every girl in the context to make sure she is sitting on a blanket.[20] Presumably the latter case, which corresponds to the underinformative sentences and the correct *all of* sentences, requires slightly more processing than the former cases, which correspond to the correct *some of* and the semantically incorrect

[20] There may be exceptions to this; Newstead (1988), for example, reviews experimental evidence showing that the meaning of *all of* may be fuzzy, particularly in the case of large sets—so that comprehenders may except *all of* when it refers to, for example, 998 entities out of a set of 1000. This suggests that there may be instances in which comprehenders verifying *all of* do not necessarily check every entity in the context, but just check until the number of entities that meet their criteria reaches some threshold which may be close to, but slightly below, the total number of entities in the context. This is unlikely to be the case in the present experiment, where the number of entities in each context is small enough to fall within the subitizing range (Degen & Tanenhaus, 2011).

sentences. The present experiments cannot yet rule out the possibility that these different quantifiers involve different types of verification strategies and that these different strategies yield different ERP signatures.

In short, then, there are potential differences across conditions in the present study that are due to quantification rather than to scalar implicature, and it is important to be cognizant of these differences. A valuable direction for future research in this area of inquiry would be to compare neural responses to pragmatic violations against other sorts of semantic violations. For instance, responses to pragmatically inconsistent *some of* in the context of an *All*-type picture could be compared to responses to different quantifiers for which the upper bound is part of the quantifier's inherent meaning, rather than a bound added through an enrichment process. This is the case, for example, for the complex quantifier *only some*, the upper bound of which is not defeasible (that is to say, unlike with *some*, it is not possible to say "Only some of the X are Y; in fact, all of the X are Y" without contradicting oneself. *Only some* is commonly used as a control quantifier in experiments on scalar implicature (see, e.g., Breheny et al., 2006; Bott et al., 2012), although a potential concern with this method is that the presence of *only* may induce additional semantic composition operations not invoked by bare *some*. Another option would be to test a bare quantifier with a similar lower-bounded meaning as *some of* and with an inherent rather than an inferred upper bound. In Mandarin, for example, the quantifier *shǎoshù –de* (少数的, "the smaller portion of" or "less than half of") may have a stronger upper-bounded meaning than *yǒu –de* (有的, "some of") and its upper-bounded meaning may be less defeasible (Jiayu Zhan, unpublished data). It should be noted, however, that the fact that participants are less tolerant of "underinformative" instances of this quantifier does not necessarily mean its upper-bounded meaning is part of the word's inherent semantics rather than an inference-based meaning (given that tolerance varies even among "pragmatic" scalars, see Doran et al., 2009), and currently there is not sufficient theoretical or empirical evidence to determine whether the upper bound of *shǎoshù –de* is qualitatively different than that of *yǒu –de*.

2.5.4.2. Does the sustained negativity reflect pragmatics or semantics?

An additional question left open in the present study is whether the operations implicated here reflect inferential pragmatic processing, or a different kind of semantic processing. While I have been referring to *some of* in the context of an *All*-type picture as "pragmatically" inconsistent, for ease of exposition, the grammatical account of scalar inference holds that the *not all* interpretation of the quantifier results from a semantic inference rather than a pragmatic one (see Section 1.1). On the basis of the present studies it is not possible to distinguish these two accounts. The fact that different ERP patterns were observed for "pragmatically" inconsistent *some of* and "semantically" inconsistent *all of* does not necessarily mean that the former process was pragmatic and the latter semantic; the difference in ERPs could be due to other factors, such as the availability of an alternate interpretation in the scalar implicature case but not in the *all of* case. In short, the mere presence of different brain responses to the different inconsistencies is not sufficient evidence to rule out the possibility that these reflect different types of semantic processing, rather than pragmatic versus semantic processing.

Nonetheless, from the present studies one can conclude that information due to scalar inference is processed differently than information inherent to the word's meaning. Above it was proposed that the difference was related to the ability to inhibit or revise the inference-based meaning. The specific nature of the inference through which this distinct meaning is realized, however, remains an open question for future research.

CHAPTER 3: THE ROLE OF CONTEXT AND PROCESSING LOAD IN SCALAR INFERENCING

3.1. INTRODUCTION[21]

The ERP studies reported in the previous chapter showed evidence that violations based on the upper-bounded, "pragmatic" meaning of a scalar quantifier are processed differently than violations based on the lower-bounded, "semantic" meaning of a scalar quantifier. This finding suggests that these different aspects of meaning are represented differently. However, those experiments do not show how the upper-bounded meaning is realized in the first place, as they tested quantifiers that mismatched with already-generated "all" or "some" representation of a picture. The goal of the experiments reported in this chapter, then, is to investigate how scalar inferences are actually realized. Rather than measuring responses to violations, these experiments adopt a violation-free design in which scalars (again the quantifier *some of*) are embedded in either contexts that encourage an inference or contexts that do not. Specifically, the experiments aim to test whether the realization of a scalar inference evokes a processing cost, an issue which is a point of fundamental disagreement among models of scalar inference processing. Background on this research question is presented in the following section.

3.2. PSYCHOLINGUISTIC INVESTIGATIONS OF SCALAR INFERENCE REALIZATION

Section 1.1 presented several competing accounts of scalar implicature processing, the most prominent among these being the classes of *default accounts* and *context-driven accounts*. As described there, these accounts make different predictions about the speed,

[21] Portions of this chapter are adapted from Politzer-Ahles & Fiorentino (in press) and Politzer-Ahles & Fiorentino (forthcoming).

context-dependency, and processing cost of inferencing. Particularly, traditional default accounts predict that inferences are realized immediately, in all contexts and without processing cost. Traditional context-driven accounts, on the other hand, predict that inferences are realized at a delay, only in certain contexts, and that the process is costly. Below, psycholinguistic studies testing each of these three predictions are reviewed. (It will be seen that these predictions are not wholly independent—that is to say, many of the studies reviewed below examine two or three of these predictions at once.)

3.2.1. Speed of scalar inferencing

Many recent studies of scalar inference processing have examined the speed at which inference-based meanings (i.e. *not all*) are realized using the visual world eye-tracking paradigm or its variant, the look-and-listen eye-tracking paradigm. Such experiments examine whether participants can use the inference-based meaning of a quantifier to rapidly restrict its possible reference. For example, in the visual world paradigm used for these studies (Huang & Snedeker, 2009, 2011; Panizza et al., 2011; Grodner et al., 2010; Degen & Tanenhaus, 2010), participants may see an array of pictures including girls and boys, and socks and soccer balls. One girl is holding all of the soccer balls present in the array, and another holding some but not all of the socks present in the array. The participant hears a sentence such as "Click on the girl who has some of the socks in the picture", and eye fixations are measured to test whether the participant looks preferentially to the appropriate referent rapidly—if preferential looking to the appropriate target emerges before the disambiguating noun *socks* is heard, and as rapidly as preferential looking triggered by semantically unambiguous quantifiers like *all* or numbers like *two or three*, this would be evidence that the participant rapidly realized the scalar inference (*some of* = *not all of*) and used it to establish appropriate reference. Several studies have indeed found this pattern

(Grodner et al., 2010; Degen & Tanenhaus, 2010), whereas others have instead found that the emergence of the inference-based interpretation was delayed relative to semantic interpretations (Huang & Snedeker, 2011, 2009; Panizza et al., 2011). Numerous design differences between the studies may contribute to the difference in results. In particular, the presence of numerals in the latter experiments but not the former ones could account for the difference, in the lack of other quantifiers in the studies that observed rapid inferencing may have allowed participants to establish a one-to-one relationship between quantifiers and referents and "pre-encode" each referent as corresponding to one quantifier or the other (Huang et al., 2010; Hartshorne & Snedeker, submitted).

Breheny, Ferguson, and Katsos (in press) have questioned the results of experiments using this paradigm, noting that according to some theories, scalar inferences within definite descriptions (*the girl that has some of the…*) are thought to be either unavailable or at least derived through different steps than typical scalar inferences. In their experiment, participants watch a video of someone pouring two different types of water (water with orange slices versus water with lime slices) into different bowls, such that all of one type of water (e.g., the water with limes) and only some of the other type (the water with oranges) is poured out. Participants then hear verbal descriptions of the video, in which the quantifiers are not embedded in definite descriptions, e.g. "The man poured <u>some of the</u> water with oranges into bowl A…". Filler items included quantifiers other than *some of* and *all of* (e.g., quantifiers such as *both of* and *a few of*), but not numerals. In this study participants looked to the correct bowl about as quickly after *some of* as they did after *all of*, suggesting that the inference was computed rapidly.

In a parallel line of research, participants have been instructed to make True/False judgments as quickly as possible, in response to underinformative statements such as "Some elephants are mammals" that have a semantic interpretation that is true and a pragmatic interpretation that is false (Noveck & Posada, 2003; Bott et al., 2004; Feeney et al., 2004; Chevallier et al., 2008; Bott et al., 2012). Participants are often shown to take longer to verify the quantifier's pragmatic interpretation (i.e., to respond "false" after realizing that it is not the case that not all elephants are mammals) than the semantic interpretation (i.e., to respond true after realizing that there are elephants that are mammals). While many of these experiments do not distinguish between the amount of time taken to realize the inference and the amount of time taken to confirm or disconfirm whether the sentence is true under that interpretation, the study by Bott and colleagues (2012) does suggest that realizing the inference itself takes time. Using a speed-accuracy tradeoff paradigm, they found that participants took longer to respond to *some* sentences (e.g. "Some elephants are mammals") than to *only some* sentences (e.g., "Only some elephants are mammals") when they were asked to interpret *some* as meaning *not all*, but did not take longer to respond to these sentences than to *at least some* sentences (e.g. "At least some elephants are mammals") when they were asked to interpret *some* as meaning *at least one*. The authors suggest that this indicates the upper-bounded interpretation of the quantifier is realized differently than the lower-bounded interpretation (consistent with the Gricean notion that the lower-bounded interpretation is inherent and semantic, whereas the upper-bounded interpretation is pragmatically added), and that realizing the upper-bounded interpretation takes additional time. On the other hand, Feeney and colleagues (2004) found a different pattern of results from the previous speeded verification studies; these authors found that participants took longer to make logical responses (i.e., based on the quantifier's lower-bounded semantic interpretation) than pragmatic responses (based on the quantifier's upper-bounded pragmatic interpretation).

In sum, evidence regarding the speed at which scalar inferences are realized remains mixed. The results of some studies suggest that inferences are realized at a delay, while others suggest that inferences are realized just as rapidly as semantic meanings.

3.2.2. Context-sensitivity of scalar inferencing

Many studies have investigated whether certain aspects of the context influence the ultimate outcome of scalar inferencing, i.e., whether a sentence is judged to have a pragmatic or a semantic reading based on the authors' introspection (e.g. Levinson, 2000; Chierchia, 2004, among others) or on experimental evidence (e.g. Geurts & Pouscoulous, 2009; Foppolo, 2007, among others). Some contextual factors that influence the ultimate realization of scalar inferences include the presence or absence of lexical alternatives in the context whether the quantifier is partitive (*some of*) or bare (*some*), whether the scalar expression has contrastive stress, whether *some of* is prosodically reduced into *summa*, and the syntactic position that the scalar expression occupies (see Section 1.1). Such investigations, while forming an important part of our understanding of the nature of scalar implicature, are not necessarily informative on the matter of the psychological realization of scalar inferences, as all the competing psychological models can account for offline judgments. As described above, context-driven models in general assume that the inference is simply not realized in such cases, and default models assume that it is realized but then cancelled through context-updating mechanisms (see Levinson, 2000, for a description of this process). Thus, this section will focus on experimental evidence probing whether context influences the initial realization of inferences.

The realization of the pragmatic interpretation of a quantifier can influence expectations about upcoming words in a sentence (Nieuwland et al., 2010; Hunt et al., 2013). This fact has been used to examine whether or not pragmatic meanings are realized in certain contexts and not others. Breheny, Katsos, and Williams (2006), examined reading times to the Greek equivalent of *the rest* in contexts that bias readers towards making the inference ("upper-bound" contexts, where what is relevant to the discourse is whether *not all* is true, and thus *some of* is likely to be interpreted as *not all of*) and in contexts that do not ("lower-bound" contexts, where what is relevant is whether *any* is true, and thus *some of* is unlikely to be interpreted as *not all*); see the examples in (5), translated from Greek.

5) a. **Upper-bound**: Mary asked John whether he intended to host all of his relatives in his tiny apartment. John replied that he intended to host *some of his relatives*. The rest would stay in a nearby hotel.

   b. **Lower-bound**: Mary was surprised to see John cleaning his apartment and she asked the reason why. John told her that he intended to host *some of his relatives*. The rest would stay in a nearby hotel.

The rest was read more quickly in upper-bound contexts, which encourage the realization of the inference, than in lower-bound contexts, which do not. The authors argue that *the rest* is more strongly expected and easier to integrate into the discourse when *some of* has been interpreted as meaning *not all of* (because this interpretation makes the reader aware that there is another subset of relatives in the discourse that has not been mentioned yet), whereas it is less strongly expected and more difficult to integrate when *some of* has not been interpreted in this way. Thus, the faster reading times in the upper-bound (inference-supporting) context indicate that the pragmatic interpretation had been realized online in that context and not in the irrelevant (inference-nonsupporting) context. (This study also examined whether context affected reading times at the quantifier itself; those results will be discussed in the next section.)

In a similar study, Hartshorne and Snedeker (submitted) also manipulated the number of words intervening between *some of* and *the rest* (and, by extension, the amount of time readers had to realize the pragmatic meaning), and found that the pragmatic meaning was realized in time to facilitate reading of *the rest* only when there was intervening lexical material between the quantifier phrase and *the rest*. This finding suggests that realizing the scalar inference is not just context-sensitive but also takes extra processing time—although it should be noted that the difference in results between the experiment with intervening material and the experiment without intervening material could be due to differences in the syntactic structure of the stimuli rather than differences in the amount of time participants had available to realize the inference.

Additional evidence regarding context-sensitivity derives from visual-world eye-tracking studies. Whereas the studies described above tested the sensitivity of scalar inferencing to local information structure (relevance of *not all* to the discourse: Breheny et al., 2006) and semantic structure (entailment polarity of the environment in which the quantifier is embedded: Hartshorne & Snedeker, submitted), preliminary evidence from Huang, Hahn, and Snedeker (2010) suggests that inferencing is also sensitive to global experimental context. These authors performed a visual world eye-tracking study similar to those described in the previous section, and between participants manipulated the presence of numbers in the filler items. For some participants, all items in the experiment used quantifiers (*some of* or *all of*), making the contrast between *some of* and *all of* more salient; for others, fillers used numbers (*two of* or *three of*), which should both reduce the salience of the contrast between *some of* and *all of*, and provide more felicitous lexical alternatives for referring to targets (that is to say, participants should be slower to interpret "the girl who is holding *some of* the balls" as referring to a girl holding two out of three of the balls, because presumably it would be more felicitous to refer to this target using a numeral). They found that looks to the target were earlier in the experimental context with only quantifiers than in the experimental context with both quantifiers and numbers, suggesting that the overall experimental context influenced the speed with which scalar inferences were realized.

Another aspect of context shown to influence the realization of scalar inferences is the epistemic state of the speaker. As described in Section 1.1, the pragmatic account of scalar inferencing assumes that inferences are realized because the hearer expects the speaker to be as informative as possible, and infers that if the speaker knew *all of* to be true then the speaker would have said *all of* rather than *some of*. (Contrast this account with the grammatical account of Chierchia and colleagues, 2012, which takes scalar inferencing to be a semantic process triggered by linguistic structure.) Crucially, in Gricean reasoning, the assumption that the speaker *knows* the stronger quantifier to have not been true is a necessary (but not sufficient) step in deriving the scalar inference. Bergen and Grodner (2012) have shown that inferences are less likely to be derived online when a scalar term is uttered by a speaker who is not fully informed. Using a self-paced reading design similar to that of Breheny and colleagues (2006), they found that reading times for *the rest* were faster in context where the implicit speaker of the sentences was assumed to have full knowledge of the referent set ("I meticulously compiled the investment report. Some of the real estate investments had lost money. The rest…") than in those where the implicit speaker had only partial knowledge ("I skimmed the investment report. Some of the real estate investments had lost money. The rest…"). Such results suggest that the online realization of scalar inferences is sensitive to speaker knowledge as well as to linguistic context. Converging results have been observed by Breheny, Ferguson, and Katsos (in press) who tested particularized conversational implicatures, rather than scalar implicatures, using a look-and-listen task, and again found that the realization of the inference was faster in situations where the speaker had full knowledge of the situation.

3.2.3. Processing cost of scalar inferencing

The issue of whether realizing the pragmatic meaning entails processing cost is also unresolved. De Neys and Schaeken (2007; see also Dieussaert et al., 2011) provide some evidence that it does: when judging underinformative sentences that were true if the quantifier was interpreted semantically and false if it was interpreted pragmatically (e.g. "Some oaks are trees"), participants were less likely to interpret the quantifier pragmatically if they were engaged in a concurrent dot memory task which burdened their executive processing resources. Studies using the speed-accuracy trade-off paradigm (Chevallier et al., 2008; Bott et al., 2012) have shown that participants are more likely to interpret a quantifier pragmatically when given more time to respond, suggesting that limiting their time to respond makes then unable to access the required processing resources in time. A limitation of studies investigating overt judgments is the difficulty of determining whether what is affected (by processing time or by concurrent processing load) is specifically inference generation, or other strategies related to evaluation and decision-making necessary for the overt response (see Huang & Snedeker, 2009; see also, however, Bott et al., 2012, for an attempt to isolate these components in an overt judgment experiment).

On the other hand, attempts to directly measure processing costs evoked by scalar inferencing have obtained mixed results. The studies described above investigating context effects on the realization of inferencing (Breheny et al., 2006; Bergen & Grodner, 2012; Hartshorne & Snedeker, submitted), in addition to testing reading times at *the rest* as an indicator of whether an inference was realized or not, also tested whether scalar terms elicit longer reading times in contexts where an inference is realized. For instance, Breheny and colleagues (2006) examined reading times to the Greek equivalent of *some of* in an upper-bounded context which encourages the comprehender to realize the inference, and in a lower-bounded context which does not (English translations of sample stimuli from their study are repeated in (6)). They hypothesized that if the realization of an inference requires processing effort, then the quantifier would be read more slowly in the upper-bound context; on the other hand, if the realization of the inference is automatic but the cancellation of the inference requires processing effort, then the quantifier would be read more slowly in the lower-bound context.

6) a. **Upper-bound**: Mary asked John whether he intended to host all of his relatives in his tiny apartment. John replied that he intended to host *some of his relatives*. The rest would stay in a nearby hotel.

   b. **Lower-bound**: Mary was surprised to see John cleaning his apartment and she asked the reason why. John told her that he intended to host *some of his relatives*. The rest would stay in a nearby hotel.

This study found longer reading times on the quantified phrase in inference-supporting contexts, suggesting that the realization of the inference was effortful. The contexts differed in more ways, however, than just whether they encouraged a scalar inference. In particular, the upper-bound context (which was meant to facilitate inferencing) mentioned the noun (*his relatives*) which was then repeated in the quantified phrase (e.g., "Mary asked John whether he intended to host all of *his relatives* in his tiny apartment. John said that he intended to host some of his relatives"), whereas the lower-bound context did not mention the noun (e.g., "Mary asked John why he was cleaning his apartment. John said that he intended to host some of his relatives"). Thus, the increased reading times could be due to the infelicity of repeating the noun (rather than using a pronoun) in that context. Indeed, in Hartshorne and Snedeker's (submitted) experiment and an eye-tracking experiment by Lewis and Phillips (2011), both using a similar design but avoiding this repeated noun effect, no difference was observed at the quantified phrase, even though the inference was realized by the time *the rest* was read.

A recent experiment by Bergen and Grodner (2012), on the other hand, included no repeated noun penalty or other confounding differences between contexts and yet found a slowdown at the quantifier itself, replicating the effect that Breheny and colleagues (2006) observed at the quantifier + noun phrase. This study used a different context manipulation: rather than manipulating the boundedness of the information-structural constraints in the context (as in Breheny et al., 2006, and Lewis & Phillips, 2011) or the entailment polarity of the environment in which the quantifier is embedded (as in Hartshorne & Snedeker, submitted), the study by Bergen and Grodner (2012) manipulated the knowledge of the implicit speaker. In the inference-supporting context, the implicit speaker had full knowledge of the situation (e.g., "I meticulously compiled the investment report; <u>some of</u> the real estate investments lost money"), whereas in the inference-nonsupporting context the speaker only had partial knowledge of the situation (e.g., "I skimmed the investment report; <u>some of</u> the real estate investments lost money").

In sum, although the results of overt judgments suggest that scalar inferencing is at least sensitive to the availability of processing resources, it is currently unclear whether the realization of a scalar inference evokes a *directly measurable* processing cost when it does occur. It should be noted that the contexts used to bias participants towards or against realizing an inference differ across these studies. Breheny and colleagues (2006) and Lewis and Phillips (2011) manipulated information structure (comparing upper-bound versus lower-bound contexts); Hartshorne and Snedeker (submitted) manipulated the entailment polarity of the semantic environment (comparing upwards entailing versus downwards entailing [conditional] environments), and Bergen and Grodner (2012) manipulated the epistemic state of the implicit speaker. These experiments also differed in the composition of their fillers, which could influence the extent to which participants are able to expect *some of* and *the rest* in the critical regions: fillers in Breheny et al. (2006), and fillers in Bergen & Grodner (2012) included (among other fillers) passages in which the inference is cancelled ("…some of the real estate investments lost money; in fact, they all did"). Only Hartshorne and Snedeker (submitted) included fillers specifically chosen to balance the number of items with and without *the rest* mentioned.

A different kind of evidence for processing costs may be found in the visual world eye-tracking study by Panizza and colleagues (2011). These authors found no evidence for rapid inferences—that is to say, looks to target remained at chance until well after the quantifier in their study, and participants only managed to preferentially fixate the target after hearing the disambiguating noun (e.g. "socks" or "soccer balls"). The authors also found that when *some of* was interpreted pragmatically, participants took longer to fixate on the target after hearing the disambiguating noun, compared to their performance in a context in which *some of* was unlikely to be interpreted pragmatically (a downward entailing semantic environment, in which the quantifier was embedded within an *if-then* statement: "If a boy has some of the paperclips, then point to him"). They argue that this may be evidence that realizing the scalar implicature occupied the participants' processing resources and prevented them from immediately using the lexical disambiguation information. This is, however, a post-hoc account based on an unexpected pattern of data which their experiment was not designed to test, motivating additional research to further explore potential processing costs.

In contrast to the above findings, several studies have suggested that it is the upper-bounded semantic meaning, rather than the lower-bounded pragmatic meaning, that requires extra effort. Feeney and colleagues (2004) found that, when reading underinformative sentences which were true when the quantifier was interpreted semantically but false when it was interpreted pragmatically, participants with higher working memory span were more likely to judge the sentences semantically (i.e., as "true"); this suggests that inhibiting the pragmatic interpretation and retrieving the semantic interpretation requires extra effort. Garrett & Harnish (2007), examining another type of pragmatic meaning (standardization implicitures) found that sentences were read more slowly in a context that cancels the implicature than in one that enables it, suggesting that the cancellation of the pragmatic interpretation and retrieval of the semantic interpretation is costly. The event-related potential (ERP) studies reported in the previous chapter are also consistent with this account; I have interpreted the sustained negativity elicited during the processing of underinformative quantifiers as an index of extra processing difficulty associated with retrieving the dispreferred semantic reading of the quantifier. The ERP studies, then, seem to suggest that it is the realization of the lower-bounded semantic reading, rather than the upper-bounded pragmatic reading, that is effortful. It should be noted, however, that the ERP studies probed responses to a quantifier that was presented after a picture was already viewed and encoded in memory; thus, retrieval of the semantic reading of the quantifier may have been difficult because the preferred pragmatic reading was already expected before the sentence was seen or heard. In the previous studies reviewed, on the other hand, quantifiers were either presented without a context (as in the case of the verification-time experiments) or with a context that presumably did not particularly bias the participant towards expecting one quantifier or the other (in the case of the reading-time experiments). Thus, whereas the reading-time experiments aimed to more directly probe the *generation* of scalar inferences, the ERP studies reported in the previous chapter aimed to probe the *revision* and processing of the meaning of *some of* after a scalar inference had already been generated.

3.2.4. Remaining questions

While several studies have found evidence that the realization of scalar inferences may be delayed and context-sensitive (although these results, particularly regarding speed, have also been challenged), few experiments have successfully linked these issues to processing costs. The majority of experiments showing slowdowns or context sensitivity have failed to show corresponding processing costs (with the possible exceptions of the studies by Bergen and Grodner, 2012, and Bott and colleagues, 2012). The lack of evidence for processing cost in many of these paradigms poses a conundrum. According to context-driven theories of scalar implicature, the reason for pragmatic meaning to be realized at a delay is precisely that the realization of this meaning is effortful and thus should be avoided when not necessary. Studies showing evidence for delays or context sensitivity in the realization of scalar inferences but failing to directly show increased processing cost raise the important question of where the slowdowns and context sensitivity come from. Is scalar inferencing generation associated with an increased processing cost that simply has not been detected yet in these paradigms? If so, what is the nature of this processing cost? The remainder of this chapter outlines two experiments that aim to address these questions.

3.3. THE PRESENT EXPERIMENTS

The experiments reported in this chapter test whether the realization of a scalar inference triggers an immediate processing cost that is directly measurable. As mentioned above, previous studies investigating this matter are equivocal. Breheny, Katsos, and Williams (2006) reported that scalar inferencing triggered a reading time slowdown, but this slowdown is likely to be due to irrelevant features of the materials used; Huang and Snedeker (submitted) found no such slowdown and Bergen and Grodner (2012) did. Furthermore, results of the ERP studies reported in the previous chapter are difficult to link directly to the question of whether inferencing is effortful, given that those studies tested how previously-realized inferences are processed and revised, rather than how such inferences are realized in the first place.

The experiments reported here adopt the design of Breheny, Katsos, and Williams (2006), but use maximally similar upper- and lower-bound contexts. A full sample set of materials is shown in (7). In this study, the only difference between the contexts is whether the context sentence uses the quantifier *all* (7a,c; compare to the upper-bound example from Breheny et al., 2006, in (5)) or *any* (7b,d).

7) a. **Upper-bound *some*:** Mary was preparing to throw a party for John's relatives. / She asked John whether *all* of them were staying in his apartment. / John said that / some of them / were. / He added / that / the rest / would be / staying / in a hotel.

b. **Lower-bound *some*:** Mary was preparing to throw a party for John's relatives. / She asked John whether *any* of them were staying in his apartment. / John said that / some of them / were. / He added / that / the rest / would be / staying / in a hotel.

c. **Upper-bound *only some*:** Mary was preparing to throw a party for John's relatives. / She asked John whether *all* of them were staying in his apartment. / John said that / only some of them / were. / He added / that / the rest / would be / staying / in a hotel.

d. **Lower-bound *only some*:** Mary was preparing to throw a party for John's relatives. / She asked John whether *any* of them were staying in his apartment. / John said that / only some of them / were. / He added / that / the rest / would be / staying / in a hotel.

Including *all* in the context makes the upper bound relevant in the discourse and thus encourages the comprehender to interpret *some of* as *not all of*, whereas *any* makes the upper bound irrelevant and discourages the inference. Importantly, unlike in the study by Breheny and colleagues (2006), this is the only difference between contexts, so reading time differences at the quantifier and quantified phrase cannot be due to a repeated noun penalty.

Furthermore, to verify whether the inference is ultimately realized, a sentence with "the rest" is included after the critical sentence with *some of*. If the reader has interpreted *some of* as meaning *not all of* (i.e., in the upper-bounded context of (7a), "Mary asked John whether *all of them* were staying in his apartment; John said that *some of them* were"), then she is aware of a remaining set of referents (e.g. "relatives") and thus more easily able to link "the rest" with a referent. On the other hand, this linking process should be more difficult when the reader has not interpreted *some of* as meaning *not all of* (i.e., in the upper-bounded context of (7b), "Mary asked John whether *any of them* were staying in his apartment; John said that *some of them* were"). Therefore, faster reading times at "the rest" in the upper-bound than the lower-bound context indicate that the inference has been realized in the upper-bound but not the lower-bound context.

"The rest" also provides a secondary test of the speed of inferencing. As mentioned above, Hartshorne and Snedeker (submitted), found faster reading times at "the rest" in the inference-supporting context when "the rest" appeared about 2500ms after the quantifier but not when it appeared about 900ms after; the authors took this as evidence that the inference takes over 900ms to realize. The current study will examine whether the inference is realized at a potentially different range of delays than those tested by Hartshorne and Snedeker (submitted), and whether the length of the delay affects the magnitude of the context effect.

In the first experiment, participants read vignettes such as those described above, without any concurrent processing load. In the second experiment the same stimulus set was used, but participants were given a concurrent processing load: listening to (and trying to ignore) irrelevant background speech (Martin, Wogalter, & Forlano, 1988). Examining how participants' interpretations of scalar quantifiers change when their processing resources are being recruited by the concurrent task makes it possible to test the role of processing resources in scalar implicature and how processing resources and context interact. While previous offline studies have suggested that the availability of processing resources influences the extent to which scalar inferences are realized overtly (De Neys & Schaeken, 2007; Dieussaert, 2011), no study has yet used online measures to examine whether processing load inferences the realization of scalar inferences during sentence processing.

Finally, participants in both experiments completed a battery of cognitive assessments, including measures of working memory span, cognitive control, pragmatic ability, and logical strategies. Several studies have suggested that individual differences in the extent to which comprehenders realize scalar inferences may be related to individual differences in more general cognitive abilities such as working memory (Feeney et al., 2004; Dieussaert et al., 2011) or pragmatic ability (Nieuwland et al., 2010). Therefore, accounting for these individual differences may provide a fuller picture of how scalar implicatures are processed across many individuals. Furthermore, investigating the relationship between scalar implicature and individual-level cognitive resources is important for its own sake, as it can provide information about the nature of the resources used for inferencing. While context-driven accounts of scalar inference processing assume that inferencing requires processing resources, these accounts have not yet articulated specifically what kinds of processing resources these might be. Thus, a variety of cognitive data were collected in this study to test whether individual differences in online, implicit realization of scalar inferences could inform our understanding of the nature of the processes underlying inference realization. The following tests were administered:

- The **Autism-Spectrum Quotient** (Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001), a 50-item questionnaire which measures the traits associated with Autism Spectrum Disorder in adults. Scores on the Communication subscale of this instrument have been shown to be correlated with participants' brain responses to manipulations of scalar implicature, and may be related to either participants' general pragmatic ability or their strategies in evaluating semantic and pragmatic meanings (Nieuwland et al., 2010).
- The **Interpersonal Reactivity Index** (Davis, 1983), a 28-item questionnaire which mainly measures individual differences in empathy but also includes items testing perspective-taking abilities. Given that Gricean accounts of scalar inferencing assume that perspective-taking and awareness of the speaker's epistemic state are relevant to inferencing (Bergen & Grodner, 2012; Breheny et al., in press), it is possible that participants with greater perspective-taking abilities may be more able to realize scalar inferences. A recent functional magnetic resonance imaging (fMRI) study has also shown that individual variation on this the Perspective-Taking subscale may be related to neural activations in processing pragmatically infelicitous *even* sentences (e.g., "He can ear even

a very loud sound"), and individual variation on the Fantasy subscale may be related to making inferences to comprehend underspecified *even* sentences (e.g., inferring that a sound is quiet based on the sentence "He can hear even that kind of sound") (Sai, Jiang, Yu, & Zhou, submitted).

- **Reading span** and **counting span** tests, which require participants to recall letters or numbers while engaging in a secondary task. They are measures of working memory capacity, which has been shown to be related to participants' inferential ability (Calvo, 2001; Feeney et al., 2004; Dieussaert et al., 2011). While there is substantial disagreement over whether the working memory resources measured by these tasks are the same as those implicated in basic syntactic processing (see Caplan & Waters, 1999), it is likely that post-syntactic processes (such as discourse integration) involve these resources, and traditional context-driven accounts of scalar inferencing would classify it as a post-syntactic process.
- **Flanker** and **Stroop** tasks. In the flanker task (Eriksen & Eriksen, 1974), participants make a response based on a central target while ignoring distracter targets to the left or right that are either congruent (**>>>>>**) or incongruent (**<<><<**) with the target. In the Stroop task (Stroop, 1935), participants name colors while ignoring the incongruous words that are printed in those colors (e.g., they say "blue" when seeing the word RED written in blue ink). Both of these tasks are considered a measures of conflict control (specifically, response inhibition), which may be accessed during the negotiation between alternative meanings of the quantifier. Neural activation in these tasks has been shown to overlap with activations in making acceptability judgments of implausible sentences (Ye & Zhou, 2009a, b); while the task in the present study is very different than that task and does not involve overt acceptability judgments, it is nevertheless possible that

participants with greater cognitive control may also be more able to make scalar inferences, particularly when burdened with a concurrent processing load (Experiment 5).

- **Truth judgments of underinformative** sentences. This task was conducted in Experiment 3, reported above, and preliminary analyses suggested that participants who gave lower truth-value ratings in this task also took more effort to switch from the upper-bounded to the lower-bounded interpretation of *some of*. The sentences used for this task were the English equivalents of the sentences used for this task in Experiment 3 (see Appendix B:).

3.4. EXPERIMENT FOUR

3.4.1. Methods

3.4.1.1. Participants

Twenty-nine native English speakers from the University of Kansas (20 women; ages 18-56, median 19) participated in the study for payment. Participants provided their written informed consent. One male participant did not return for the second session of the experiment, in which the individual differences measures were collected, and thus that participant was included in the group analysis of reaction times but not in the individual differences analysis. One female participant was unable to complete the reading span task because of an equipment failure.

3.4.1.2. Materials

Forty-eight sets of four-sentence vignettes were constructed following the template in (7) above, repeated as (8) for convenience. Slashes indicate how the vignettes were divided into segments for the self-paced reading task (see Procedure).

8) a. **Upper-bound *some*:** Mary was preparing to throw a party for John's relatives. / She asked John whether *all* of them were staying in his apartment. / John said that / some of them / were. / He added / that / the rest / would be / staying / in a hotel.

b. **Lower-bound *some*:** Mary was preparing to throw a party for John's relatives. / She asked John whether *any* of them were staying in his apartment. / John said that / some of them / were. / He added / that / the rest / would be / staying / in a hotel.

c. **Upper-bound *only some*:** Mary was preparing to throw a party for John's relatives. / She asked John whether *all* of them were staying in his apartment. / John said that / only some of them / were. / He added / that / the rest / would be / staying / in a hotel.

d. **Lower-bound *only some*:** Mary was preparing to throw a party for John's relatives. / She asked John whether *any* of them were staying in his apartment. / John said that / only some of them / were. / He added / that / the rest / would be / staying / in a hotel.

In each set, the first sentence establishes a set of items or people (e.g., John's relatives). The second sentence establishes an upper- or lower-bound context by asking about either *all of* them or *any of* them. The third sentence includes a response to the previous indirect question, using *some of*, which is predicted to be interpreted as *not all* in the upper-bound (since "all" is relevant in that context, but was not used) but not the lower-bound context (since "all" is not relevant in that context). Finally, the fourth sentence includes a mention of *the rest* of the set. "The rest" was always followed by "would be" and two or three more segments of one or more words each. The only difference between contexts is the use of *all* or *any* in the second sentence. A full list of critical, filler, and practice stimuli, including comprehension questions (see Procedure) is included in Appendix C:.

In addition to the boundedness of the context, the quantificational expression in the third sentence was also manipulated. Each of the vignette types above also has a counterpart written using *only some of* rather than *some of* (see (8c-d)), serving to make the *not all* interpretation semantically explicit (see Minai & Fiorentino, 2010, for a discussion of the semantics of *only some*). This is important because comparing reading times between sentences in which the quantifier was interpreted pragmatically and those in which the quantifier was interpreted semantically involves comparing across sentences with different meanings, which may take different amounts of time or effort to interpret or verify (see Bott et al., 2012; Hartshorne & Snedeker, submitted). For example, evaluating whether *not all* is true may involve a different sort of reasoning than evaluating whether *at least one*, or *all*, or *none*, is true; these differences are not necessarily based on pragmatic inference. Furthermore, in the present study a lack of facilitation in reading times for *the rest* might be due to a failure to generate the pragmatic reading, or to a failure to use that information to predict and integrate upcoming words in the sentence. The goal of the present study is to examine pragmatic processing, rather than quantificational, truth-verificational, or predictive/integrative processing, and thus it is important to include a control comparison to isolate those factors from factors relating to pragmatic inferencing. If a difference between upper- and lower-bound conditions is due to pragmatic inferencing rather than other factors, then that difference should appear in the implicit upper-bound (*some of*) sentences but not in the explicit upper-bound (*only some of*) sentences.

In addition to the critical stimuli, 144 filler vignettes were created. Forty-eight follow the same format as the critical sentences but do not include *the rest*; this is both to make sure participants cannot predict *the rest* in every item and to make sure that *some of* is not always associated with *the rest* (which is an explicit cue to the inference). Forty-eight use *all of* rather than *some of* or *only some of* in the third sentence, to make sure participants cannot predict *some* *of* or *only some of* in every item; these items also do not include *the rest*. The last 48 use various other quantifiers in the third sentence (*many of*, *most of*, *several of*, *a few of*, *none of*, and numbers) to increase the variety of lexical alternatives to *some of* present in the experimental context, which has been shown to influence the speed and outcome of scalar inferencing (Degen & Tanenhaus, 2011).

3.4.1.3. Procedure

Participants were tested individually, each in two one-hour sessions. Participants completed the self-paced reading task in the first session, and the individual difference measures in the second session. The seven individual difference tasks were administered in a random order for each participant.

3.4.1.3.1. Self-paced reading

Participants read the vignettes in a non-cumulative moving-window self-paced reading paradigm (Just et al., 1982), administered using the Presentation software package (Neurobehavioral Systems, Inc.). In each trial, the passage was shown on the screen with all the characters replaced with dashes; the participant pressed a button on a gamepad to show a phrase (at which point the dashes were replaced with the phrase). With each button press, the currently displayed phrase turned back into dashes and the next phrase was displayed. Line breaks always occurred after the first context sentence, the second context sentence, and "he/she added" in the final sentence, as shown in (9):

9) Mary was preparing to throw a party for John's relatives. /
She asked John whether all of them were staying in his apartment. /
John said that / some of them / were. / He added /
that / the rest / would be / staying / in a hotel.

This ensured that the critical segments (*some of them* and *the rest*) never appeared adjacent to a line break.

Participants were instructed to read the sentences for comprehension at a natural reading speed. One-third of the sentences were followed by comprehension questions, e.g. "Who was Mary throwing a party for?" The comprehension questions never targeted aspects of the passage that depend upon the interpretation of quantifiers. The main experiment was preceded by eight practice items. The procedure took 40-50 minutes to complete, with five breaks.

3.4.1.3.2. Flanker task

The flanker task was administered using the Presentation software package (Neurobehavioral Systems, Inc.). Stimuli consisted of rows of one or five angle brackets (>, <). There were six types of stimulus, based on direction of the target (facing left or right) and type of flankers (no flankers [e.g. **<**], congruent flankers [e.g. **<<<<<**], or incongruent flankers [e.g. **>><>>**]). A 600x300px light gray rectangle remained on the screen throughout the task, and stimuli were presented at the center of it in 30pt Times New Roman font. Each trial began with a fixation point (+) presented in the center of the rectangle for a random duration between 500 and 1750 ms, followed by the stimulus; the target bracket appeared in the same spot as the fixation point. The participant's task was to press, as quickly as possible, the button (left or right shift key) corresponding to the direction the target bracket was pointing. The stimuli remained on screen until the participant's response or for 1500 ms. If the participant responded incorrectly or did not respond within the allotted time, a feedback message ("Wrong!" or "Too slow!") was presented in red at the center of the screen for 500ms. The inter-trial interval was 1500 ms. Participants performed three blocks consisting of 48 fully randomized trials each; the task was preceded by a practice block of 24 fully randomized trials.

3.4.1.3.3. Stroop task

The paper-based version of the Stroop task described by Hinkin, Castellon, Hardy, Granholm, and Siegle (1999) was used. In this task, participants saw lists of 100 items and were instructed to read them aloud one item at a time (the instructions emphasized reading as quickly as possible without making a mistake); the time taken to complete each list was measured using a stopwatch. Participants completed three lists, always in the same order: the first list consisted of uppercase color words printed in black, which they had to read aloud; the second consisted of "XXXX"s printed in different colors of ink, the colors of which they had to name; and the third consisted of uppercase color words printed in incongruous colors of ink, the colors of which they had to name. The four colors used (both for words and ink) were red, blue, green, and yellow. The full list of stimuli for the Stroop task is included in Appendix D:.

3.4.1.3.4. Counting span

Participants completed a computer-mediated version of the counting span task described by Engle, Tuholski, Laughlin, and Conway (1999; see also Conway, Kane, Bunting, Hambrick, Wilhelm, & Engle, 2005). The task was administered using Paradigm (Perception Research Systems, Inc.). Participants saw 15 items, each consisting of two to six trials. On each trial, the participant saw an array of three to nine blue dots, one to nine blue squares, and one to nine green dots. The particpant's task on each trial to count aloud the number of blue dots and then repeat the final count, after which the next trial was presented. After completing all two to six trials in an item, the participant was asked to recall the final counts for that item in order. Items were presented in the same order for all participants. Within an item, no two trials had the same number of blue dots. Before beginning the test, participants completed a practice block consisting of three two-trial items.

3.4.1.3.5. Reading span

Participants completed a computer-mediated version of the reading span task described by Kane, Hambrick, Tuholski, Wilhelm, Payne, and Engle (2004; see also Conway, Kane, Bunting, Hambrick, Wilhelm, & Engle, 2005). The task was administered using Paradigm (Perception Research Systems, Inc.). Participants saw 12 items, each consisting of two to five trials. On each trial, the participant saw a visually-presented sentence followed by a "?" and a capital letter. Sentences were either semantically anomalous (e.g., "During the week of final spaghetti, I felt like I was losing my mind") or semantically acceptable (e.g., "During the winter you can get a room at the beach for a very low rate"). The participants' task was to read the sentence aloud and then make a semantic acceptability judgment using the mouse. After making the judgment, the participant was to say the letter aloud, after which the next trial was presented. After completing all two to six trials in an item, the participant was asked to recall the final letters of each trial in that item, in order. Items were presented in the same order for all participants. Within an item, no two trials had the same letter following the sentence. Before beginning the test, participants completed a practice block consisting of three two-trial items.

3.4.1.3.6. Autism-Spectrum Quotient, Interpersonal Reactivity Index, and truth-value ratings of underinformative sentences

Each of these instruments was written in Hypertext Markup Language (HTML) and administered at a computer using Perl CGI. Participants were shown all items at once in a list format, and chose the appropriate rating for each item by selecting a radio button with the mouse.

3.4.1.4. Data analysis

3.4.1.4.1. Reaction time data

Reading times for filler items and for the first two segments of the critical items (the context segments which were presented as entire sentences) were excluded from all analyses. The remaining reading times were log-transformed for normality, and outliers for each participant and item removed based on visual inspection.[22] Linear mixed models with crossed random intercepts for participants and items (Baayen et al., 2008) were fit with predictors Quantifier (*some*, *only some*), Boundedness (*upper, lower*), and sentence Segment, and model comparison was conducted with log-likelihood tests.[23] Accuracy was analyzed using generalized linear mixed models with predictors Quantifier and Boundedness. Evaluation of the significance of model coefficients was conducted using Markov Chain Monte Carlo sampling.

3.4.1.4.2. Flanker task

Incorrect responses were removed from the analysis, and outliers for each participant removed based on visual inspection. Reaction times were transformed using the reflected reciprocal transformation (each observation was divided by 1 and then subtracted from the

[22] This method is recommended by Baayen (2008: 266). The pattern of results reported below was also observed using other outlier-trimming methods such as a flat criterion (as in Hartshorne & Snedeker, submitted), a subject-wise standard deviation criterion, and a hybrid method based on that described in Breheny et al., 2006 (first removing observations below 150ms or greater than 3 times the overall mean of observations in a given region; then removing observations that differ by more than three standard deviations from that subject's mean for that region).

[23] A common practice in both ANOVA-based and mixed-model-based self-paced reading research is to test separate models for each segment of the stimuli. In the experiments reported here, results were instead based on tests of a single model with Segment as a factor (a method used by, for example, Grodner, Gibson, and Tunstall, 2002). This was done both in order to minimize the chance of observing spurious effects at some segments due to conducting multiple comparisons, and provide a stronger test of whether crucial effects (i.e., those at "some of" and "the rest") were limited to those segments, rather than being general effects emerging because of the different contexts.

highest raw reaction time in the data set) to yield an approximately normal distribution. For each participant, the transformed reaction times were regressed on Target Direction (left or right), Fixation Duration (the duration for which the fixation point was displayed before the presentation of the target) and Flanker Type (Congruent, Incongruent, or None). Each participant's coefficient for Flanker Type == Incongruent represented the flanker effect—the amount by which the participant slowed down when responding to incongruent flankers as compared to congruent flankers. The other regressors were nuisance regressors to reduce error. Descriptive statistics for the flanker effects are given in Appendix E:.

3.4.1.4.3. Stroop task

For each participant, the time taken to complete the Color Naming list was subtracted from the time taken to complete the Incongruent list to represent the Stroop effect—the amount by which the participant slowed down when naming colors that were incongruent with their background as compared to colors with neutral backgrounds. Stroop effect scores were log-transformed to approximate a normal distribution. Descriptive statistics for the Stroop effects are given in Appendix E:.

3.4.1.4.4. Counting span and reading span

Each participant's performance on the recall portion of each span task was scored according to the *partial-credit unit scoring* procedure described by Conway and colleagues (2005). In this procedure, each item gets a score reflecting how many what proportion of trials the participant recalled correctly in that item (e.g., a participant correctly recalling 2 trials out of 5 would receive a score of .4 for that item) and the scores of the 15 items are then averaged, yielding an aggregate score between 0 and 1 for each participant, with higher scores reflecting greater recall accuracy.

Each participant's accuracy on the secondary processing task task (reading or counting) was also calculated as the proportion of trials with correct performance. Finally, recall and processing scores were converted to z-scores, and a composite working memory score was calculated for each participant by averaging the recall z-score and the processing z-score.[24] The analyses reported below were all conducted using the composite scores. One participant did not participate in the reading span task; the group mean was substituted for this participant's score. Prior to analysis, the composite scores for each span task were reflected, square root transformed, and re-reflected to approximate a normal distribution. Descriptive statistics for the working memory span tasks are given in Appendix E:.

3.4.1.4.5. Autism-Spectrum Quotient

Participants were scored according to the guidelines given by Baron-Cohen and colleagues (2001). In the Autism-Spectrum Quotient, half the items are designed such that an "agree" answer corresponds to an Autism-like trait, and half are designed such that a "disagree" answer corresponds to an Autism-like trait. Each participant receives a total score (between 0 and 50) which is the number of items to which she gave an answer that corresponds to an abnormal or Autism-like behavior. Furthermore, each of the 50 items is associated with one of

[24] Conway and colleagues (2005) suggest using only recall scores in computing the working memory score, and not considering scores on the processing task. Waters and Caplan (1996), however, argue that composite scores which take into account both the recall and the processing components of a task should be used. First of all, correlations between the recall and processing scores on working memory span tasks tend to be positive but small (Waters & Caplan, 1996; Kane, 2004), suggesting that the processing scores contain information not reflected in the recall scores; this was also the case for the present dataset (Reading Span: $r = .22$, $p = .116$; Counting Span: $r = .35$, $p = .009$). Secondly, Waters and Caplan (1996) found that composite scores, compared to recall scores, showed better test-retest reliability and stronger correlations with other reading comprehension and memory tasks.

five subscales (Social Skill, Attention Switching, Attention to Detail, Communication, and Imagination), so the participant also receives five subscale scores, each between 0 and 10. Scores on the Social Skill, Attention Switching, and Imagination subscale were log-transformed, and scores on the Communication subscale were square root transformed. Descriptive statistics for the subscales are given in Appendix E:.

3.4.1.4.6. Interpersonal Reactivity Index

Participants were scored according to the guidelines given by Davis (1983). In the Interpersonal Reactivity Index, half the items are designed such that an answer of "this statement describes me very well" corresponds to a high value on the corresponding scale (e.g., high perspective-taking ability, high empathy, etc.), and half are designed such that a "this statement does not describe me very well" answer corresponds to high value. Each item receives a score of 0 to 5 points (because the participant's responses are on a 5-point Likert scale). Each participant receives a total score (between 0 and 112, since the test consists of 28 items) with a higher score corresponding to overall higher interpersonal/empathetic ability. Furthermore, each of the 28 items is associated with one of four subscales (Perspective-Taking, Fantasy, Empathetic Concern, and Personal Distress), so the participant also receives four subscale scores, each between 0 and 28. Scores on the Fantasy and Empathy subscales were reflected, square root transformed, and re-reflected to approximate a normal distribution. Descriptive statistics for the subscales are given in Appendix E:.

3.4.1.4.7. Truth-value ratings of underinformative sentences

Participants' mean truth-value and mean naturalness ratings for the underinformative sentences in the task were recorded; ratings for the true sentences were not analyzed. Descriptive statistics for the rating task are given in Appendix E:.

3.4.1.4.8. Regression with individual difference measures

Because these tasks were administered in a random order for each participant, it was necessary to test whether scores on any tests were substantially influenced by the order in which the test appeared in a session, to rule out potential fatigue effects. Scores on none of the tests were significantly correlated with the order in which the test occurred during the session (flanker: *r* = .167, *p* = .237; Stroop: *r* = -.005, *p* = .97; count span: *r* = .018, *p* = .9; reading span: *r* = -.064, *p* = .654; Autism-Spectrum Quotient total score: *r* = -.012, *p* = .929; Interpersonal Reactivity Index total score: *r* = -.202, *p* = .148; truth ratings of underinformative sentences: *r* = .083, *p* = .556).

	Stroop	AQ Social Skill	AQ Attention Switching	AQ Attention to Detail	AQ Communication	AQ Imagination	IRI Perspective	IRI Fantasy	IRI Empathy	IRI Distress	Truth Rating	Reading Span	Counting Span
Stroop **(1.42)**													
AQ Social Skill **(1.85)**	-0.01												
AQ Attention Switching **(1.29)**	0.05	0.06											
AQ Attention to Detail **(1.32)**	0.17	0.15	0.11										
AQ Communication **(2.01)**	0.14	0.51	0.3	0.01									
AQ Imagination **(2.44)**	0.24	0.36	0.03	0.11	0.41								
IRI Perspective **(1.67)**	0.13	0.04	-0.00	0.1	0.01	-0.05							
IRI Fantasy **(1.58)**	-0.22	0.09	0.12	0.17	-0.02	-0.38	0.07						
IRI Empathy **(2.27)**	0.15	-0.24	0.07	-0.07	0.04	-0.37	0.37	0.14					
IRI Distress **(1.50)**	0.06	-0.02	0.3	-0.05	0.28	-0.03	-0.28	0.15	0.22				
Truth Rating **(1.15)**	0.19	-0.06	0.14	-0.03	0.03	-0.1	-0.1	-0.17	0.04	0.02			
Reading Span **(1.37)**	-0.23	0.11	-0.03	0.16	-0.14	-0.23	-0.02	0.3	-0.05	-0.04	-0.03		
Counting Span **(1.66)**	-0.07	-0.1	-0.01	0.16	-0.15	-0.17	0.05	0.24	-0.22	-0.11	0.01	0.36	
Flanker **(1.35)**	-0.05	0.14	0.07	0.26	-0.15	-0.12	0.14	0.04	-0.01	-0.08	0.1	0.21	-0.05

Table 5. Correlation matrix of individual difference variables in the model, with corresponding variance inflation factors (VIFs).

The data from each measure were sphered (such that each measure had a mean of 0 and a standard deviation of 1), and the total scores on the Autism-Spectrum Quotient and Interpersonal

Activity Index were excluded from the analysis.[25] Table 5 shows the correlation matrix of the variables that were kept for further analyses, as well as the variance inflation factors (VIFs) for each variable.

3.4.2. Results

3.4.2.1. Accuracy

Participants responded correctly to 94% of items in the upper-bound *some* condition, 89.7% in lower-bound *some*, 94.8% in upper-bound *only some*, and 91.4% in lower-bound *only some*. There were no significant differences in accuracy across conditions (Quantifier: $\chi^2 = 0.29$, $p = .591$; Boundedness: $\chi^2 = 2.59$, $p = .107$) and no interaction ($\chi^2 < 0.01$, $p = .968$).

3.4.2.2. Reading times: group analysis

Figure 16 shows the reading times for the last two sentences of the vignettes. It is evident that, for *some* sentences, "the rest" was read more slowly in the lower-bounded context, whereas such a pattern was not observed in *only some* sentences. It is also apparent that there is no slowdown at the quantifier in *some* sentences in the upper-bound context. Statistical analysis confirmed these observations.

[25] Including the total scores as well as the subscale scores would make the set of measures highly multicollinear, since the total scores on each test are a mathematical combination (a sum) of the subscale scores. Even with sphering, the condition number of the entire set with these variables included 7.939012×10^{16}, and the variance inflation factors (VIFs) for all the Autism-Spectrum Quotient and Interpersonal Reactivity Index measures were infinite. A condition number of 30 or more is considered to indicate a high level of multicollinearity (Baayen, Feldman, & Schreuder, 2006). With a set of predictors that is highly multicollinear it becomes difficult or impossible to estimate the effect of any given predictor (Baayen et al., 2006). Removing the Autism-Spectrum Quotient and Interpersonal Reactivity Index total scores from the analysis was sufficient to reduce the condition number of the sphered data to 4.51, which is considered low, and to reduce the VIFs such that they were all within the acceptable range (below 4).

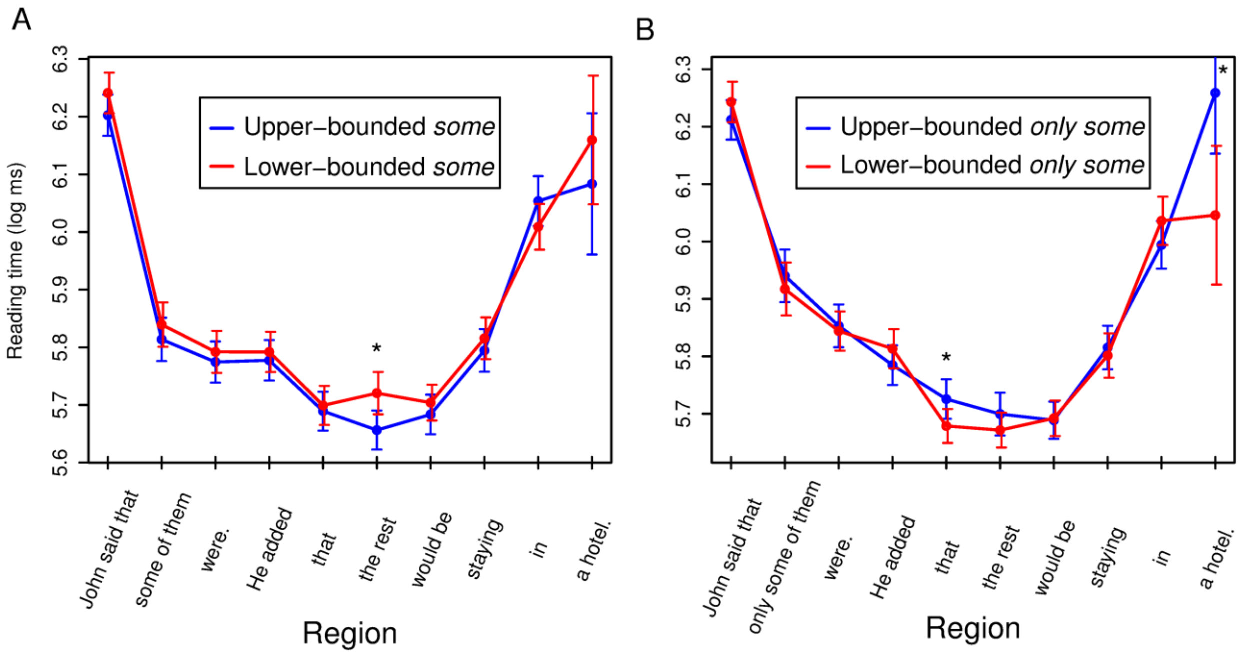


Figure 16. Reading times by segment for the last two sentences in *some* vignettes (panel A) and *only some* vignettes (panel B). Segments showing a significant effect of boundedness for a given quantifier type are indicated with an asterisk. Error bars represent ±2 standard errors of the mean.

After outlier removal (see Data analysis), 12,543 observations remained for analysis. There was a significant three-way interaction between Region, Quantifier, and Boundedness ($\chi^2(9) = 26.18$, $p = .002$).[26] For *some* sentences, reading times for "the rest" were significantly slower in the lower-bound than upper-bound context ($b = 0.068$, SE = 0.022, $t = 3.11$, $p = .002$); a marginal pattern in the same direction was also observed in the following segment ($b = 0.036$, SE = 0.022, $t = 1.66$, $p = .096$). No significant difference was observed at "the rest" in *only some* sentences, and the trend was in the opposite direction ($b = -0.026$, SE = 0.022, $t = -1.17$, $p = .242$). The only segments where *only some* sentences showed a boundedness effect were "that" (the region preceding "the rest"; $b = -0.046$, SE = 0.022, $t = -2.08$, $p = .038$) and the last two segments ($b = 0.038$, SE = 0.022, $t = 1.73$, $p = .083$; $b = -0.163$, SE = 0.022, $t = -2.95$, p

[26] Standard deviations for the random effects in this model were as follows: Items: 0.034; Participants: 0.17; Residual: 0.288.

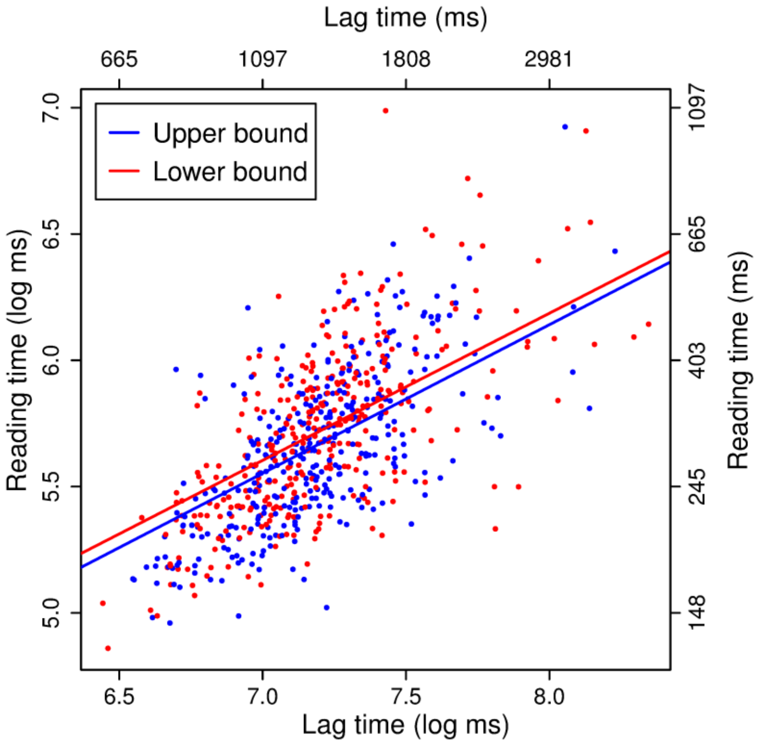


Figure 17. Relationship between reading times on "the rest" and lag between the quantifier and "the rest" for upper-bound (blue) and lower-bound (red) contexts. Points represent individual observations, and regression lines represent predictions from a mixed model with fixed effects of Boundedness, Lag Time, and their [non-significant] interaction. The bottom and left axes show log lag time and log reading time respectively, and the top and right axes show raw lag time and raw reading time.

= .003).[27] No significant effect of context was observed at the quantifier ("some of" or "only some of") or the following two regions, either for *some* sentences (*b*s < 0.031, SEs = 0.022, *t*s < 1.41, *p*s > .156) or for *only some* sentences (|*b*|s < 0.02, SEs = 0.022, |*t*|s < 0.9, *p*s > .368).

Because Hartshorne and Snedeker (submitted) found an effect of context when "the rest" appeared about 2500ms after the quantifier but not when it appeared about 900ms after, the lag between quantifier and "the rest" in the implicit upper-bound (*some of*) vignettes was calculated.

[27] These segments represent the 11th and 12th segments. It should be noted, however, that the stimuli differed in length: for 40 stimuli the 11th segment was the last in the vignette, whereas for eight stimuli the 12th was the last. Therefore, reading times for these segments are not particularly meaningful, given that the 12th segment represents a very small number of items, and the 11th represents a mixture of final and non-final segments.

The average lag was 1440 ms. A mixed model on the reading times at "the rest" and the following region, for the *some* sentences only, showed that the effect of context did not interact with the lag time ($\chi^2(2) = 0.24$, $p = .627$), thus not providing evidence that the effect of context on scalar inferencing emerged only at long lag times. As illustrated in Figure 17, the effect of context (at "the rest") remains the same regardless of the lag time.

3.4.2.3. Reading times: individual differences analysis

The goal for the analysis of individual differences in reaction times was to identify whether any of the individual measures collected predicts inferencing ability specifically. Therefore, the focus of the analysis will be on individual measures which were involved in, at the least, three-way interactions with Segment and Quantifier.[28] An interaction with Quantifier is necessary to show that the measure is related to inferencing in particular, rather than other aspects of evaluating upper- versus lower-bounded meanings; and an interaction with Segment is necessary to show that the relationship between reading times and the individual differences measure is limited to segments of the vignette that are expected to show effects of inferencing (i.e., "some of" and "the rest"), rather than being a general effect throughout reading. Individual difference measures that also interact with Boundedness would be of particular interest, since the focus of this experiment was to test whether the effect of Boundedness in the critical segments of *some of* sentences would be moderated by individual difference measures. Presumably, however, moderation of entirely context-independent aspects of inferencing by individual differences could also manifest as a three-way interaction of Segment, Quantifier, and an individual

[28] Interactions between individual differences measures were not tested. That is to say, these measures each were allowed to interact with the within-participant factors Segment, Quantifier, and Boundedness, but not with the other between-participant measures.

difference measure. (This would be the case if, for instance, if reading times for critical segments were modulated by individual differences in *some of* sentences but not *only some of* sentences.)

The omnibus model revealed no significant four-way interactions, but did reveal a significant three-way interaction between Segment, Quantifier, and IRI-Fantasy ($\chi^2(7) = 18.63$, $p = .009$) and a marginal three-way interaction between Segment, Quantifier, and AQ-Social ($\chi^2(7) = 12.48$, $p = .086$). These interactions were resolved by Segment.

The Quantifier × IRI-Fantasy interaction was marginal at the segment *before* the quantifier ($\chi^2(1) = 2.74$, $p = .098$); at this segment, IRI-Fantasy had a non-significant positive association with reading times for *only some of* sentences ($b = 0.07$, SE = 0.06, $t = 1.32$, $p = .109$), but had little effect for *some of* sentences ($b = 0.04$, SE = 0.06, $t = 0.60$, $p = .431$). At the quantifier itself ("only some of them" or "some of them"), the Quantifier × AQ-Social interaction was significant ($\chi^2(1) = 7.13$, $p = .008$); at this segment, reading times for "only some of them" decreased somewhat as AQ-Social subscale scores increased ($b = -0.10$, SE = 0.07, $t = -1.31$, $p = .097$), but this was not the case for reading times for "some of them" ($b = -0.02$, SE = 0.07, $t = -0.32$, $p = .679$). At the following segment, the Quantifier × IRI-Fantasy interaction reached significance ($\chi^2(1) = 6.49$, $p = .011$); here, reading times in *some of* sentences increased as a function of IRI-Fantasy ($b = 0.09$, SE = 0.06, $t = 1.66$, $p = .055$), but reading times for *only some of* sentences were relatively unaffected ($b = 0.03$, SE = 0.06, $t = 0.52$, $p = .543$). Two segments later (at "that", the third segment after the quantifier, and the first segment before "the rest"), both interactions were marginal (Quantifier × AQ-Social: $\chi^2(1) = 3.03$, $p = .082$; Quantifier × IRI-Fantasy: $\chi^2(1) = 2.93$, $p = .087$). At this segment, AQ-Social did not have a significant effect in either type of sentence (*only some* sentences: $b = 0.03$, SE = 0.0523567, $t = 0.63$, $p = .458$; *some* sentences: $b = -0.01$, SE = 0.05, $t = -0.10$, $p = .906$); on the other hand, IRI-Fantasy had a significant positive effect in *only some* sentences ($b = 0.09$, SE = 0.0498854, $t = 1.81$, $p = .033$), and an even larger positive effect in *some* sentences (*b* = 0.13, SE = =0.05, *t* = 2.57, *p* = .030). Finally, The Quantifier × IRI-Fantasy interaction was marginal again in the segment following "the rest", "would be" ($\chi^2(1) = 3.28$, *p* = .070); again, IRI-Fantasy had a larger positive effect on reading times in *some of* sentences (*b* = 0.09, SE = 0.05, *t* = 1.82, *p* = .030) than *only some of* sentences (*b* = 0.05, SE = 0.05, *t* = 1.06, *p* = .188).

3.4.2.3.1. Exploratory individual differences analysis

As mentioned above, the analysis did not find evidence for individual differences in context-specific inferencing costs (which would have required four-way interactions between any individual difference measure, Segment, Quantifier, and Boundedness). Given that the analysis with 28 participants may have lacked power to find such a four-way interaction, an exploratory analysis was conducted to further examine potential individual differences in context effects. For this analysis, each participant's context effect (log reading time for lower-bounded items minus log reading time for upper-bounded items) was calculated for each segment in both *some* and *only some* sentences, and these context effects were submitted to a mixed model as described above. Interactions were found between Segment, Quantifier, and the following individual predictors: Stroop ($\chi^2(7) = 14.80$, *p* = .039), AQ-Social ($\chi^2(7) = 12.46$, *p* = .086), and AQ-Imagination ($\chi^2(7) = 13.73$, *p* = .056). These interactions were resolved by Segment.

At the segment containing the quantifier ("only some of them" or "some of them"), the Quantifier × AQ-Social interaction reached significance ($\chi^2(1) = 4.85$, *p* = .028); reading times for "only some of them" were faster in the upper-bound than lower-bound context on average, but the effect decreased as AQ-Social subscale scores increased (*b* = -0.07, SE = 0.04, *t* = -1.571, *p* = .140); the context effect for "some of them", however, was not substantially affected by this predictor (*b* = 0.03, SE = 0.04, *t* = 0.60, *p* =.542). At the following segment, all three interations

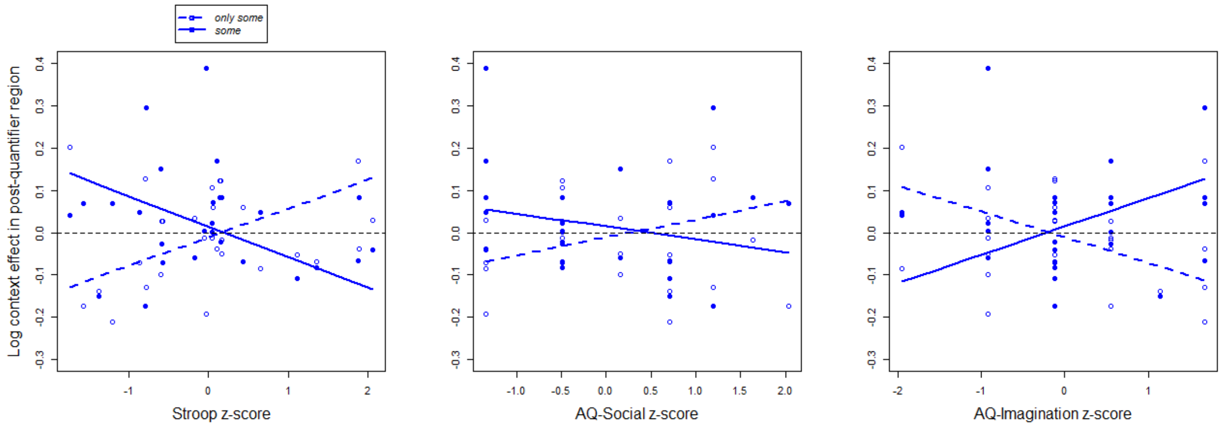


Figure 18. Context effects at the segment following the quantifier. The thin black dashed line indicates 0 (no context effect).

were significant or marginal (Quantifier × Stroop: $\chi^2(1) = 13.22$, *p* < .001; Quantifier × AQ-Social: $\chi^2(1) = 3.55$, *p* = .060; Quantifier × AQ-Imagination: $\chi^2(1) = 11.42$, *p* = .001). These interactions are illustrated in Figure 18. The Quantifier × Stroop interaction indicated that for *only some* sentences the effect of context tended to be negative (that is, faster reading times in lower-bounded than upper-bounded sentences) for participants with lower Stroop effects (greater cognitive control), and positive for participants with higher Stroop effects (*b* = 0.07, SE = 0.04, *t* = 1.80, *p* = .068); whereas for *some* sentences this pattern was reversed (*b* = -0.07, SE, 0.04, *t* = -1.92, *p* = .068). The Quantifier × AQ-Social interaction was of a similar nature (*only some*: *b* = 0.04, SE = 0.04, *t* = 1.071, *p* = .291; *some*: *b* = -0.03, SE = 0.04, *t* = -0.77, *p* = .443), whereas the Quantifier × AQ-Imagination interaction showed the opposite pattern (*only some*: *b* = -0.06, SE = 0.04, *t* = -1.63, *p* = .112; *some*: *b* = 0.07, SE = 0.04, *t* = 1.80, *p* = .087).

At the segment containing "the rest", the Quantifier × AQ-Social interaction was significant ($\chi^2(1) = 7.01$, *p* = .008), as was the Quantifier × AQ-Imagination interaction ($\chi^2(1) =$ 6.04, *p* = .014). As shown in Figure 19, these interactions were driven by individual differences in the processing of *only some* sentences; the context effect in *some* sentences was relatively unaffected. The Quantifier × AQ-Social interaction remained significant into the following

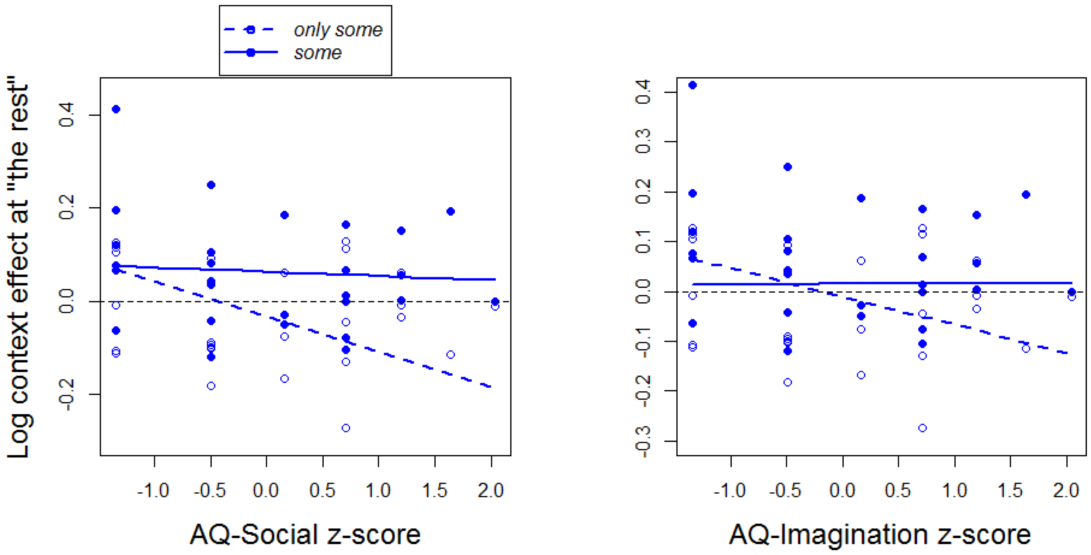


Figure 19. Context effects at "the rest". The thin black dashed line indicates 0 (no context context effect).

segment ($\chi^2(1) = 5.60$, $p = .018$), and was marginal in the second segment after "the rest" ($\chi^2(1) = 3.00$, $p = .083$). The interactions at these two segments are shown in Figure 20. While there was a slight numerical trend for *some of* sentences to show a greater context effect (slower reading times in lower-bounded contexts) in participants with higher AQ-Social subscale scores, the effects of AQ-Subscale did not reach significance in either segment, either for *some* or *only some* sentences ("the rest"+1, *only some*: $b = -0.03$, SE = 0.03, $t = -0.85$, $p = .393$; "the rest"+1, *some*: $b = 0.03$, SE = 0.03, $t = 0.82$, $p = .410$; "the rest"+2, *only some*: $b = -0.03$, SE = 0.03 $t = -0.86$, $p = .400$; "the rest"+2, *some*: $b = 0.02$, SE = 0.03, $t = 0.59$, $p = .561$.).

3.4.2.3.2. Summary of individual differences analyses

While several predictors emerged as significant in the individual differences analyses, only a few of them are likely to be related to the inferencing process itself. In the omnibus analysis, scores on the Fantasy subscale of the Interpersonal Reactivity Index modulated reading times in the spillover region after "some of them"; in the segments immediately before and after

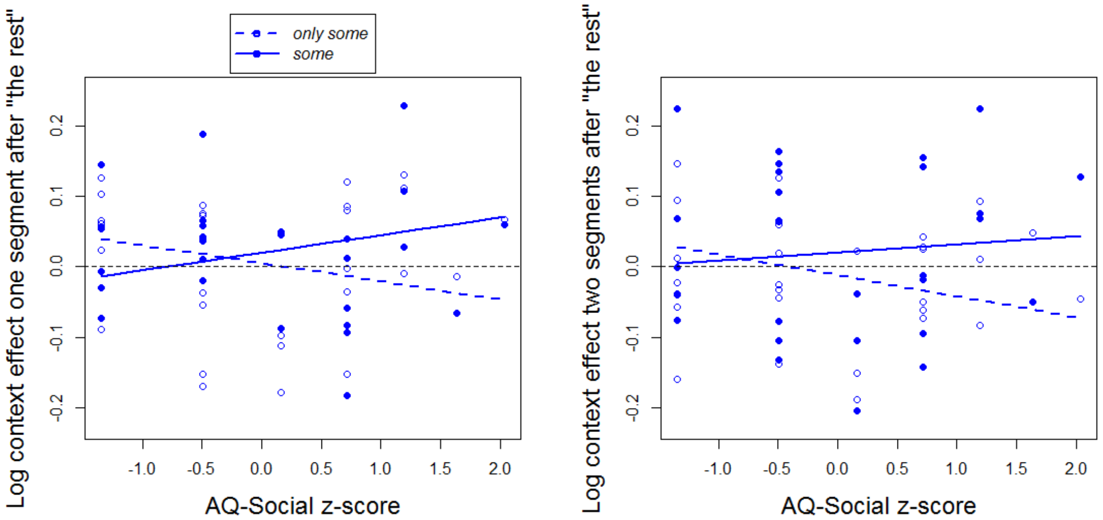


Figure 20. Context effects at "the rest". The thin black dashed line indicates 0 (no context context effect).

"the rest", scores on the same subscale modulated reading times in *some of* sentences more than reading times in *only some of* sentences. In the exploratory analysis, the context effect (difference in reading times between lower-bounded and upper-bounded sentences) at the spillover region following "some of" was modulated by cognitive control abilities as assessed by the Stroop task, and by scores on the Social Skill and Imagination subscales of the Autism-Spectrum Quotient. Potential interpretations of these effects will be addressed in the discussion (Section 3.4.3). Other predictors had effects on *only some* sentences but not *some* sentences; these effects are potentially interesting, given that comprehending *only some* may require complex syntactic composition (see Minai & Fiorentino, 2010), but they are unlikely to be related to scalar inferencing and thus are not discussed here.

3.4.3. Discussion

The results of the present experiment suggest that the ultimate realization of scalar inferences is sensitive to context—the inference is more likely to be realized in the upper-bound than lower-bound context, as evidenced by the fact that "the rest" was read faster in the former context. These results are consistent with the majority of previous studies exploiting this paradigm. Crucially, however, no evidence was found for increased processing costs—either in the form of reading time slowdowns or reading time moderation by individual differences—at the point of the scalar quantifier in the context that encourages inferencing, even using a rather liberal analysis. This finding contrasts with the results of Breheny, Katsos, and Williams (2006), who found a reading time slowdown at the quantifier and argued that the realization of a scalar inference is effortful. The results of this study, along with similar recent studies that have failed to find reading time slowdowns in similar designs (Lewis & Phillips, 2011; Hartshorne & Snedeker, submitted) suggest that the slowdown observed in that study was due to properties of the stimuli other than the pragmatic manipulation—for example, the repeated noun penalty. On the other hand, Bergen and Grodner (2012) did observe a reading time slowdown at the quantifier in a design similar to this, and without the repeated name confound; further discussion of the differences between that study and the present experiment is in the general discussion of the self-paced reading experiments below.

3.4.3.1. The facilitation effect at "the rest"

There are at least two potential interpretations of the effect observed at the mention of the complement set ("the rest"). The account made by Breheny and colleagues (2006), and in the predictions given above, is that the increased reading times in the lower-bounded context relative to the upper-bounded context reflect difficulty in integration of the word "the rest" into the discourse when the participant has not already made the scalar inference which makes her aware of the complement set. In other words, the reading times reflect the reader's trying to find a set of referents to which they can link "the rest". I will refer to this account as the *discourse linking* *account*. An alternative explanation is that the reading time slowdown in this segment in fact reflects the realization of the scalar inference in the lower-bounded context. Under this explanation, in the upper-bounded context the inference is realized immediately and effortlessly at the quantifier; this occurs because the nature of the context, combined with a reader's lexical knowledge about the scalar nature of *some of*, constitute a strong cue for making the inference. On the other hand, in the lower-bounded context, the quantifier is not a sufficiently strong cue for the inference, and thus the reader does not make the inference until she reaches "the rest", which explicitly indicates that the quantifier in this item is inconsistent with *all*; in this case, the realization of the inference may be more effortful than it was in the upper-bounded context because the cue is less effective, or because the cue comes later in the sentence (several words after the quantifier) and thus the reader needs to revise an initial interpretation. I will refer to this account as the *late inferencing account*. At present, it is not clear whether the discourse linking account and the late inferencing account are distinguishable. The discourse linking account assumes that the reading time slowdown is based on the identification of a complement set; if the meaning of quantifiers is represented in terms of sets, this may well be the same process as interpreting *some* as *not all*.[29] Furthermore, while the late inferencing account assumes that mention of the complement set triggers a scalar inference, it should be noted that the complement set could be recognized even without making the inference. Interpretation of the complement set only requires recognizing that *some* was referring to less than all, and the semantic interpretation of the quantifier (*at least one*) can be consistent with this meaning. In other words, while integrating "the rest" with the discourse requires identifying a complement set, it is not

[29] The semantic representation of quantifiers is under debate. The idea that quantifiers represent relations between sets (functions from properties to [functions from properties to truth-values]: $\langle \langle e,t\rangle, \langle \langle e,t\rangle, t\rangle \rangle$) is a central tenet of Generalized Quantifier Theory (see Barwise & Cooper, 1981). For further discussion on Generalized Quantifier Theory and alternatives, see, among others, Hackl (2009).

necessarily the case that identifying a complement set requires making a scalar inference—the scalar inference might facilitate the identification of a complement set but not be a necessary condition.

3.4.3.2. The speed of inferencing

Regarding the speed of inferencing, the results of this experiment were not wholly consistent with the conclusions drawn by Hartshorne and Snedeker (submitted) regarding the time necessary to realize an inference. In the study they report, the upper-bounded context facilitated reading times for "the rest" when "the rest" occurred about 2500 ms after the quantifier, but not when it occurred about 900 ms after the quantifier. They took this as evidence that the realization of the inference takes at least 900 ms. In the present study, however, the relationship between facilitation at "the rest" and lag time between the quantifier and "the rest" was directly tested, and the tests did not reveal evidence that the context effect only emerged at a long lag; rather, the context effect remained the same across lag times, even though the range of lag times observed in the present experiment ranged from below 900 ms to above 2500 ms. The failure to replicate this interaction suggests that the lag-time effect of Hartshorne and Snedeker (submitted) may be due to other structural properties of the stimuli and not to the speed of inference. In particular, the long-lag conditions in that experiment included adverbial phrases after the quantifier, whereas the short-lag conditions did not, as shown in (10):

10) a. **Long lag:** Addison ate some of the cookies before breakfast this morning, and the rest are on the table.

b. **Short lag:** Addison ate some of the cookies, and the rest are on the table.

This manipulation may have introduced several differences between the conditions other than just a difference in the time available to complete the inference. For instance, the lack of adverbial detail in the short-lag condition may have reduced the felicity of the more specific upper-bounded description, thus creating a global experimental context that discourages participants from computing such readings (compare to Nieuwland et al., 2010). It should be noted, however, that whereas the experiment by Hartshorne and Snedeker (submitted) manipulated the lag between the quantifier and the mention of the complement set ("the rest"), the present experiment did not manipulate that lag time, but rather used lag time variations introduced by the participants themselves as a result in their own variations in reading speed. Thus, while the present experiment did not show evidence that participants or items with more time in between the quantifier and "the rest" were more able to realize inferences, factors that contribute to the increased lag time could also be factors that inhibit inferencing. For example, if a participant has more time between the quantifier and "the rest", that participant may have relatively poor reading comprehension or working memory, and thus even though that participant has more time to realize inferences, she may also have fewer processing resources available to do so. Thus, the results of the present study together with those of Hartshorne and Snedeker (submitted) suggest that this type of inference may be realized within about 1440 ms (the mean lag time observed in the present experiment) in general, but that further research is needed to determine just how quickly the effect of inferencing emerges in this research paradigm, and to better understand whether the strength of this effect is modulated by lag time.

3.4.3.3. Individual differences in inferencing

The individual differences analysis revealed that some individual-level cognitive factors may be associated with inferencing. Reading times after *some of* increased as a function of scores on the Fantasy subscale of the Interpersonal Reactivity Index—people with higher scores spent longer reading these segments. This was not the case in *only some of* sentences. In the present study there was not an *a priori* prediction regarding this pattern and thus conclusions about this effect need to be verified through further study, but one possible interpretation of the effect is that it is due to uncertainty introduced by the ambiguous quantifier *some of*. Evidence for this account comes from an fMRI study by Sai and colleagues, in which the authors found that this subscale correlated positively with BOLD activations in the medial prefrontal cortex (mPFC) for underspecified sentences. More specifically, participants in this study read *even* sentences in Chinese which were either specified (e.g., "He can even hear such a quiet sound"), or underspecified (e.g., "He can even hear that kind of sound"); in the former case, the sentence specifies that the sound is quiet, whereas in the latter case, the listener must infer that the sound is quiet. Participants with higher scores on the Fantasy subscale showed a greater difference in activation between underspecified and specified sentence in the mPFC, a region potentially associated with mentalizing and with making inferences under uncertain situations (Jenkins & Mitchell, 2010). In the present study, participants with higher fantasizing ability may have committed more resources to considering alternative interpretations (i.e., both the semantic and the pragmatic interpretations) of *some of*. The fact that this effect did not interact with Boundedness may suggest that such participants consider both interpretations regardless of the context, a pattern not predicted by context-driven accounts of inferencing. It should be noted that this account of the effect does not explain why the effect re-emerged after "the rest", a point at which the uncertainty may have been removed (the phrase "the rest" makes the existence of the complement set explicit).

An exploratory analysis of the individual difference measures also suggested that the context effect after reading the quantifier (i.e., how much slower this segment was read in lower-bounded contexts than in upper-bounded contexts) was modulated by cognitive control (as measured by the Stroop task) and by scores on the Imagination and Social Skill subscales of the

Autism-Spectrum quotient. Recall that context-driven accounts predict faster reading times in the lower-bounded context at this point (in the terms used here, this would be a negative context effect. In fact, such an effect only emerged in a subset of participants—participants with poorer cognitive control (higher Stroop scores) and worse social skill (higher scores on the Social Skill subscale), were more likely to show a negative effect of context. This might suggest that the inferencing process was more costly for such participants, whereas for others it was relatively effortless; such an account remains to be directly tested in future study. As for imagining abilities, inspection of Figure 18 suggests that worse imagining ability (higher scores on the subscale) was associated with more *positive* context effects (which are not predicted by context-driven accounts), but participants with better than average imagining ability did not necessarily have more negative effects. A traditional prediction regarding the default account of inferencing (see Breheny et al., 2006) is that scalars in lower-bounded contexts might take longer to read because the reader must cancel the inference (see Section 3.6, General Discussion, for further discussion of this assumption). Thus, this effect might suggest that cancelling inferences is only costly for a subset of participants for whom the process of imagining other possible interpretations is difficult. This hypothesis remains to be tested in future experimentation.

3.4.3.4. An alternative means of testing for processing costs

Although the present experiment did not find evidence that the realization of scalar inferences involves an increased processing cost, it remains possible that there was a processing cost that self-paced reading times and individual differences in the skills measured are simply not sensitive to. Thus, the following experiment uses another method to test for processing costs: a dual-task design. De Neys and Shaeken (2007) and Dieussaert and colleagues (2011) found evidence that participants make fewer pragmatic responses to underinformative sentences when they are engaging in a concurrent spatial working memory task; these results were taken as evidence that inferencing requires processing resources, although it is also possible that their task influenced participants' ability to make off-line evaluations or verifications or pragmatic readings rather than to generate those readings. Thus, the following experiment adopts a concurrent task along with self-paced reading of the types of vignettes used in the previous experiment, in order to test whether concurrent task load modulates implicit realization of scalar inferences as measured by self-paced reading times. Rather than using the dot recall task used in those studies, the present study instead used a task in which participants listen to irrelevant background speech, following the design of Martin, Wogalter, & Forlano (1988). As the presence of unattended background noise has been shown to modulate sentence comprehension (Martin et al., 1988), and as it is a secondary task that is performed continuously while the self-paced reading is under way (rather than memorization before and recall after the reading task), this task may offer a greater chance of detecting an effect, compared to the dot recall task. If scalar inferencing is dependent on the availability of processing resources, then the facilitation effect at "the rest" observed in the previous experiment should be eliminated in the presence of the secondary task.

3.5. EXPERIMENT FIVE

3.5.1. Methods

3.5.1.1. Participants

Thirty-seven native English speakers from the University of Kansas (28 women; ages 18-32, median 20) participated in the study for payment. Participants provided their written informed consent. The computer failed to record data from the flanker task for one female participant. One additional female participant participated in the experiment, but her data were not used because a scripting error on the experimenter's part caused the quantifiers in the self-paced reading task not to display. Data from one female participant were removed from analysis because this participant responded with less than 75% accuracy to comprehension questions on critical trials; thus, the total number of participants included in data analysis was thirty-six.

3.5.1.2. Materials

The materials for the self-paced reading task were identical to those in Experiment 4.

The materials for the unattended listening task were two lists of words: one comprising real words, and one pseudowords. (Martin and colleagues, 1988, found that unattended real words disrupt sentence comprehension to a greater extent than unattended pseudowords, presumably because the presence of lexical information makes the parser automatically devote processing resources to recognizing real words.) The real-word list consisted of 800 English words pseudorandomly chosen from a convenience sample of texts. The novel-word list consisted of 791 pseudowords that followed English phonotactics but did not match the pronunciation of any existing English word. Two hundred eleven of these were created by changing and/or transposing several phones from words in the real-world list; 49 were novel compound stimuli from Fiorentino, Politzer-Ahles, Popescu, & Popescu (2011); 444 were novel compound stimuli from Fiorentino, Politzer-Ahles, & Pak (2012), and 87 were from another experiment in progress in our laboratory. The full list of words is available in Appendix F:.

The words were read aloud by four native speakers of English (two male and two female) who were naïve to the purposes of the study. Each participant read the real-word list first and the novel-word list second. A different random order of words in each list was used for each participant. The recording was carried out within an anechoic chamber at the University of

Kansas, using an ElectroVoice 767 microphone and a Marantz PMD-671 digital solid-state recorder sampling at 22050 Hz and in mono format. Offline processing of the recordings was conducted using Praat (Boersma & Weenik, 2012). Pauses between words were removed, all four lists were intensity-normalized, and male and female lists were combined (overlain on top of one another) to create four lists: male/real (9.5 minutes), male/novel (10 minutes), female/real (9.5 minutes), and female/novel (10 minutes). In cases where one speaker's list was shorter than another speaker's list because of faster speaking rate, a few extra tokens from the middle of that speaker's list were appended to the end of the same speaker's list to ensure that for the entirety of the sound file there were always two speakers audible. The reason for creating multi-talker lists was to reduce the salience of list intonation, which otherwise may have influenced reading times by making the self-paced reading participants synchronize their button-pressing to the rhythm of the background speech.

3.5.1.3. Procedure and data analysis

The procedure for this experiment was the same as that for Experiment 4, except that participants in the self-paced reading task also listened to the background speech over binaural headphones. On each trial, speech from one of the lists began to play at the start of the trial and continued through the end of the trial. The next time speech from the same list was to be played, it began at whatever point in the list had been reached when the last trial on that list ended. If the end of a list was reached, playback for that list began again at the start of the list. The real- and novel-word background speech conditions were randomly mixed on a trial-by-trial basis.

The procedures for all individual differences measures were the same as in Experiment 4. Data analysis for reading time data, accuracy data, and individual differences measures was all the same as in Experiment 4. One participant did not participate in the flanker task; the group mean was substituted for this participant's score. One participant each in the Stroop and flanker tasks showed effects that were more than 4 standard deviations lower than the group's mean (these participants had faster naming times in the incongruent than congruent Stroop conditions, and faster reaction times in the incongruent than congruent flanker conditions, respectively); these participants' scores were replaced with the group minima prior to sphering. One participant on the Reading Span task had a composite score nearly 4 standard deviations below the group's mean (this participant only recalled 8% of items correctly, and also misjudged the acceptability of the sentences 20% of the time); this participant's reading span composite score was replaced with with a the group minimum prior to sphering.[30]

3.5.2. Results

3.5.2.1. Accuracy

[30] Some individual difference measures had significant interactions with reading times and other predictors of interest when these correctional measures were not taken, but no longer had significant interactions after the data were treated in this way. This suggests that those significant interactions were driven by outliers.

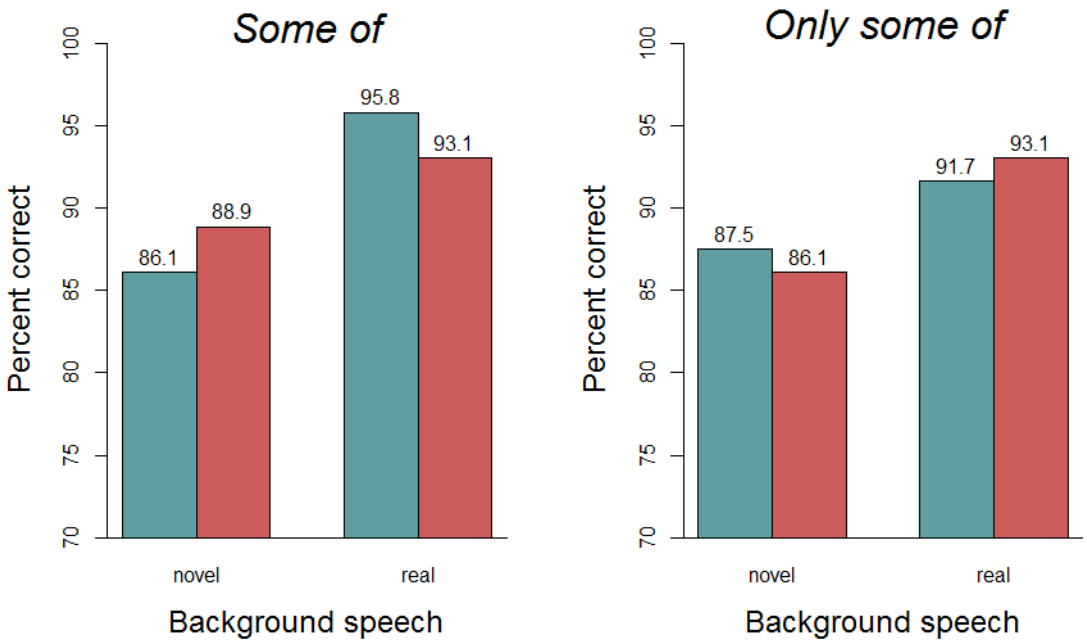


Figure 21. Comprehension accuracy on critical items in Experiment 5.

All participants performed at an average accuracy of 85% or higher. Accuracy for each condition is shown in Figure 21. A generalized linear mixed model on accuracy revealed only a main effect of Background Condition ($\chi^2(1) = 7.55$, $p = .006$), reflecting the fact that participants were more accurate on items with real-word background speech than novel-word background speech.

3.5.2.2. Reading times: group analysis

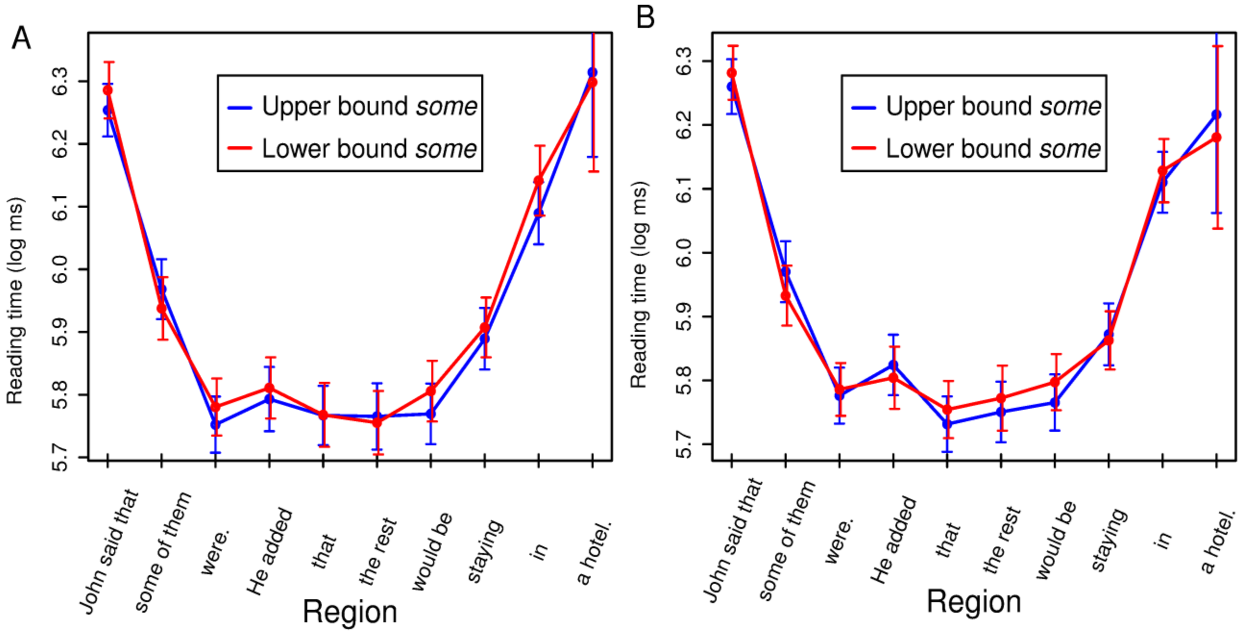


Figure 22. Self-paced reading times for *some of* sentences with real-word background speech (panel A) and novel-word background speech (panel B). Error bars represent ±2 standard errors of the mean.

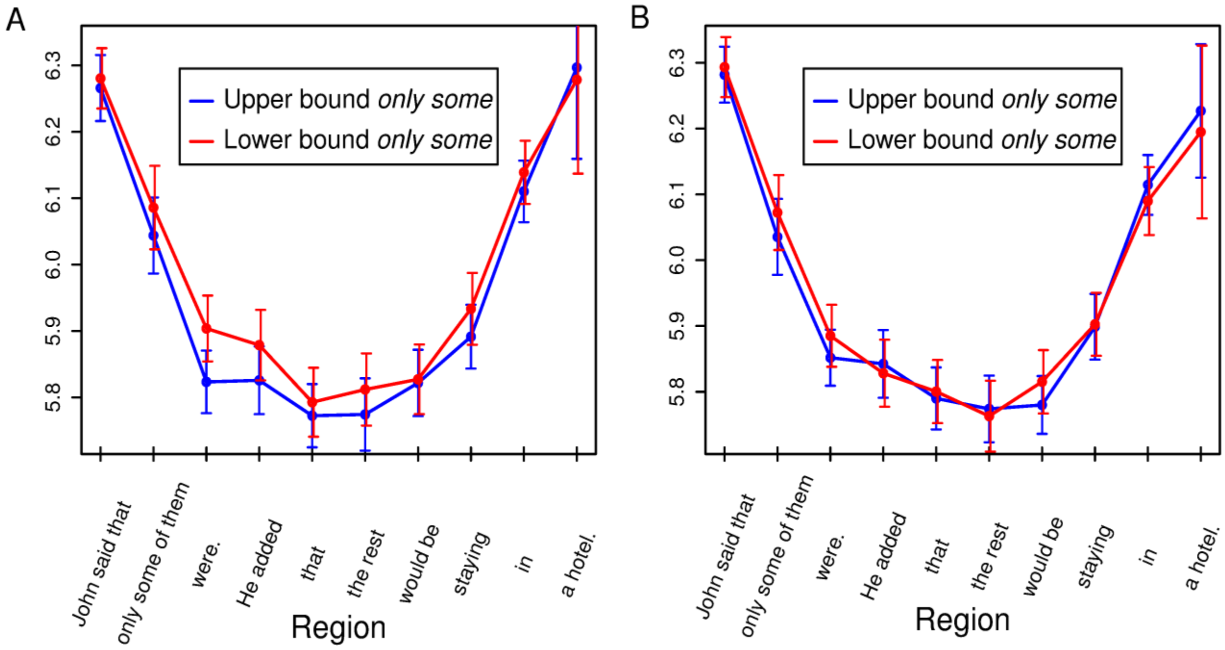


Figure 23. Self-paced reading times for *only some of* sentences with real-word background speech (panel A) and novel-word background speech (panel B). Error bars represent ±2 standard errors of the mean.

After removal of outliers, 15,166 observations remained for analysis. Reading times are shown in Figure 22 and Figure 23. Compared to the previous experiment, the difference between reading times for lower-bound and upper-bound contexts in *some of* sentences appears smaller. The quantifier in *some of* sentences also appeared to be read slightly slower in the upper-bound, implicature-supporting context. The following statistical analyses, however, demonstrate that none of these differences was significant.

Unlike in the previous experiment, the interaction between Segment, Quantifier, and Boundedness was not significant ($\chi^2(9) = 9.71, p = .375$), nor was the four-way interaction with Background Condition ($\chi^2(9) = 2.60, p = .978$).[31] These findings indicate that that there was no effect of Boundedness specific to the critical regions.[32] The other effect of interest is a marginal interaction between Background Condition and Boundedness ($\chi^2(1) = 3.56, p = .059$), reflecting the fact that overall reading times in upper-bounded sentences were not significantly affected by the lexicality of the background speech ($b = -0.01$, SE = 0.02, $t = -0.80, p = .437$), but overall reading times in lower-bounded sentences were somewhat slower with real word background speech than novel word background speech ($b = 0.02$, SE = 0.01, $t = 1.88, p = .061$).

[31] Standard deviations for random effects in the model were as follows: Items: 0.03; Participants: 0.21; Residual: 0.28.

[32] There was a significant main effect of Boundedness ($\chi^2(1) = 12.83, p < .001$), indicating that lower-bounded sentences were read more slowly overall ($b = 0.02$, SE < 0.01, $t = 3.58, p = .001$). There was also a marginal Quantifier × Boundedness interaction ($\chi^2(1) = 3.00, p = .083$). Because these effects did not interact with Segment, they do not provide any evidence for effects of context specific to the critical regions (either slowdowns for "some of" or facilitation for "the rest" when an inference is realized).

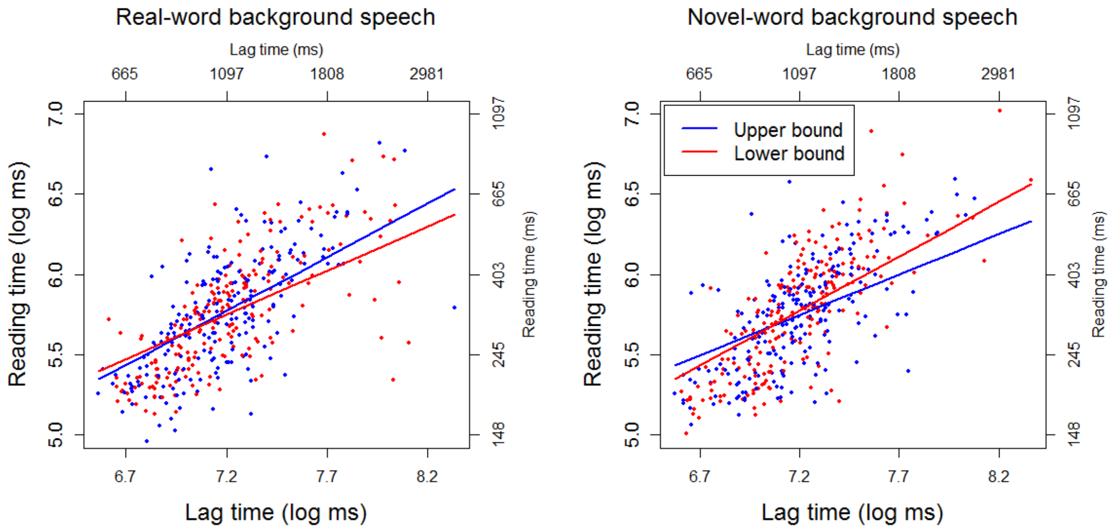


Figure 24. Relationship between reading times on "the rest" and lag between the quantifier and "the rest" for upper-bound (blue) and lower-bound (red) contexts, in trials with real-word background speech (left) and novel-word background speech (right). Points represent individual observations, and regression lines represent predictions from a mixed model with fixed effects of Boundedness, Lag Time, and their interaction. The bottom and left axes show log lag time and log reading time respectively, and the top and right axes show raw lag time and raw reading time.

As in Experiment 4, tests were conducted to examine whether the lag time between "some of" and "the rest" influenced the facilitation effect at "the rest" and the following segment. This time, a significant four-way interaction between Segment, Lag Time, Boundedness, and Background Condition did emerge in *some of* sentences ($\chi^2(1) = 5.20$, $p = .023$). Resolving the interaction by Region revealed that the effect of lag time was not moderated by Boundness or Background Condition in the region following "the rest" (χ^2s$(1) < 2.40$, *p*s > .301), whereas at "the rest" itself there was a three-way interaction between Lag Time, Boundedness, and Background Condition ($\chi^2(1) = 6.34$, $p = .012$). Resolving that interaction by Background Condition revealed that the effect of lag time was not significantly moderated by Boundedness in trials with real-word background speech ($\chi^2(1) = 2.62$, $p = .105$), but it was in trials with novel-word background speech ($\chi^2(1) = 6.57$, $p = .010$). As shown in Figure 24, when the background speech consisted of novel words, longer lag time was associated with a more positive effect of Boundedness (i.e., reading times for lower-bounded sentences became more and more slower than those for upper-bounded sentences), similar to the pattern reported by Hartshorne and Snedeker (submitted). When the background speech consisted of real words, on the other hand, the effect was in the opposite direction and was not significant

3.5.2.3. Reading times: Individual differences analysis

The analysis of individual differences in reading times was conducted following the same standards as in Experiment 4. No interactions of interest reached significance.

3.5.2.3.1. Exploratory individual differences analysis

As in Experiment 4, an exploratory analysis of the context effects was conducted. For this analysis, each participant's context effect (log reading time for lower-bounded items minus log reading time for upper-bounded items) was calculated for each segment in both *some* and *only*

Effect	**$\chi^2(7)$**	***p***	**Significant segments**
Segment × Quantifier × AQ-Imagination	20.41	.005**	5*, 6*
Segment × Quantifier × IRI-PerspectiveTaking	19.64	.006*	8***
Segment × Quantifier × CountSpan	19.62	.006*	4*
Segment × Quantifier × Background Condition × IRI-PerspectiveTaking	17.49	.014*	5*, 8*

Table 6. Interactions of interest that reached significance in the omnibus analysis. **p* < .05 ***p* < .005, ****p* < .001. The "Significant segments" column indicates the segments at which the lower-order interaction was significant. 4: the quantifier; 5: the verb following the quantifier; 6: the beginning of the following sentence ("He/she added"); 8: "the rest".

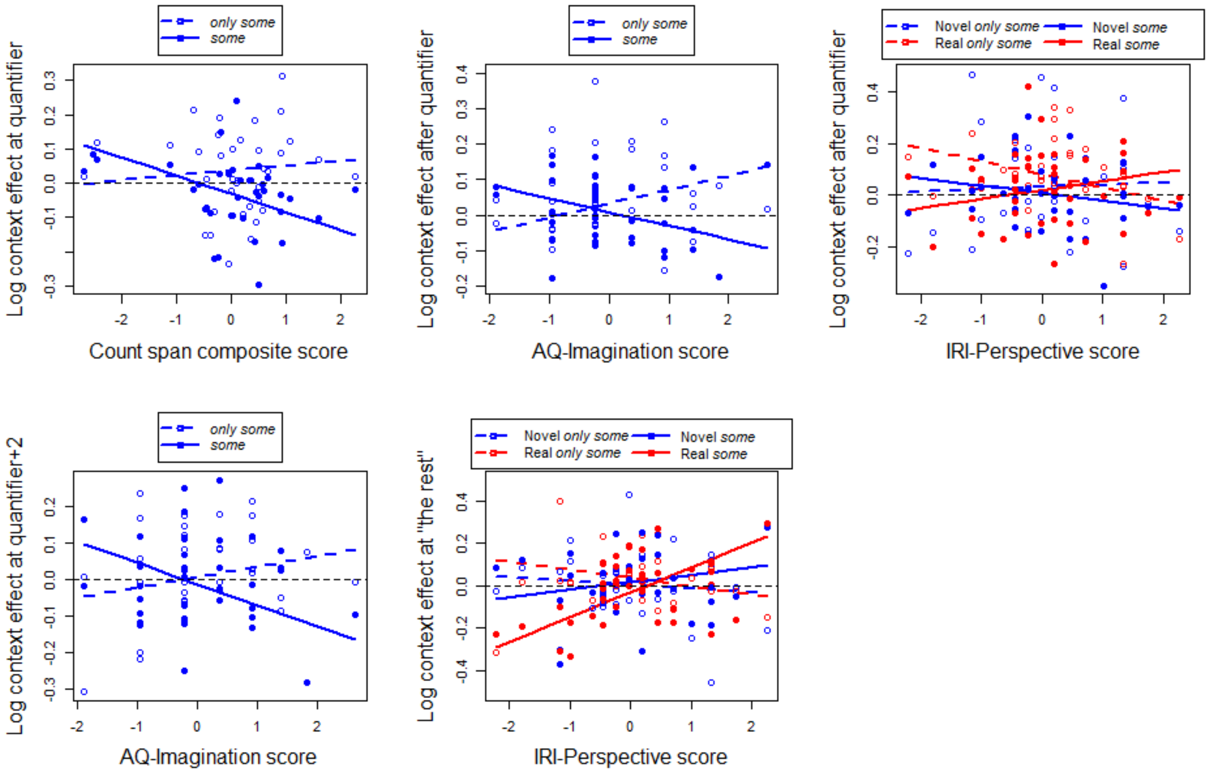


Figure 25. Relationship between individual difference measures and context effects at segments where significant interactions were observed. The *y*-axis for each subplot indicates which segment is being shown. See text for details.

some sentences, and these context effects were submitted to a mixed model as described above. Table 6 shows which individual difference measures showed significant interactions with the predictors of interest in this analysis.

At the quantifier itself, the Quantifier × Count Span interaction was significant ($\chi^2(1) = 6.60$, *p* = .010). As shown in the upper left portion of Figure 25, the context effect at the quantifier in *only some of* sentences was relatively unaffected by Count Span (*b* = 0.015, SE = 0.03, 0.50, *p* = .613); in *some of* sentences, however, negative context effects (faster reading times in lower-bound than upper-bound sentences) were marginally more likely to appear in participants with higher composite Count Span scores (*b* = -0.053, SE = 0.03, *t* = -1.83, *p* = .072).

At the following segment, the Quantifier × AQ-Imagination interaction reached significance ($\chi^2(1) = 6.74$, $p = .009$) The effects of AQ-Imagination in *only some of* and *some of* sentences were opposite, but did not reach significance in either sentence type. (*only some*: $b = 0.040$, SE = 0.03, $t = 1.19$, $p = .239$; *some*: $b = -0.039$, SE = 0.03, $t = -1.16$, $p = .255$). The Quantifier × Background Condition × IRI-PerspectiveTaking interaction also reached significance at this segment ($\chi^2(1) = 4.24$, $p =.040$), but when resolving the interaction the Background Condition × IRI-PerspectiveTaking effect did not reach significance for either *only some of* sentences ($\chi^2(1) = 1.76$, $p = .185$) or *some of* sentences ($\chi^2(1) = 2.58$, $p = .108$). Resolving the interaction by Background Condition rather than Quantifier revealed a significant Quantifier × IRI-Perspective interaction for trials with real-word background speech ($\chi^2(1) = 5.02$, $p = .025$), but not trials with novel-word background speech ($\chi^2(1) = 0.68$, $p = .409$). As shown in the upper right portion of Figure 25, IRI-PerspectiveTaking had opposite effects in *only some of* and *some of* sentences in the real-word background speech condition, but neither reached significance (*only some*: $b = -0.050$, SE = 0.03, $t = -1.47$, $p = .153$; *some*: $b = 0.035$, SE = 0.03, $t = 1.02$, $p = .305$).

Two segments after the quantifier, the Quantifier × AQ-Imagination interaction remained significant ($\chi^2(1) = 7.22$, $p = .007$). As shown in the lower right portion of Figure 25, the context effect at the quantifier in *only some of* sentences was relatively unaffected by AQ-Imagination scores ($b = 0.028$, SE = 0.04, $t = 0.80$, $p = .435$); in *some of* sentences, however, negative context effects (faster reading times in lower-bound than upper-bound sentences) were marginally more likely to appear in participants with higher AQ-Imagination scores (and thus poorer imagining ability) ($b = -0.056$, SE = 0.04, $t = -1.63$, $p = .098$).

At "the rest", the Quantifier × IRI-PerspectiveTaking interaction reached significance ($\chi^2(1) = 14.91$, $p < .001$). The Quantifier × Background Condition × IRI-PerspectiveTaking interaction was again significant ($\chi^2(1) = 3.93$, $p = .047$), indicating that the above effect was mainly driven by a significant Quantifer × IRI-PerspectiveTaking interaction for the real-word background condition ($\chi^2(1) = 21.06$, $p < .001$), whereas the Quantifier × IRI-PerspectiveTaking interaction did not reach significance in the novel-word background condition ($\chi^2(1) = 2.17$, $p = .140$). As shown in the lower-right portion of Figure 25, the context effect in *only some of* sentences was not significantly moderated by IRI-PerspectiveTaking scores ($b = -0.042$, SE = 0.03, $t = -1.45$, $p = .153$), but the context effect in *some of* sentences was ($b = 0.061$, SE = 0.03, $t = 2.10$, $p = .037$): participants with greater perspective-taking ability were more likely to show a positive effect of context (longer reading times in lower-bounded than upper-bounded contexts), and this interaction was strongest in trials with real-word background speech.

3.5.2.3.2. Summary of individual differences analyses

While the omnibus individual difference analysis revealed no significant effects, the exploratory analysis uncovered several effects that may be related to theories of inferencing. At the quantifier, working memory as measured by Counting Span scores interacted with the context effect for *some of* sentences, such that only participants with high working memory showed longer reading times for "some of" in upper-bounded contexts (the effect predicted by context-driven accounts). The Imagination subscale of the Autism-Spectrum Quotient modulated context effects in a similar way, albeit later: two segments after "some of", participants with poor imagining ability tended to show longer reading times in upper-bounded contexts. Finally, at "the rest", the context effect observed in the previous experiment (longer reading times in lower-bounded than upper bounded *some of* contexts) emerged mainly in participants with high Perspective-Taking ability (as measured on the Interpersonal Reactivity Index), rather than other participants.

3.5.3. Discussion

This experiment tested whether the presence of a concurrent processing load would modulate participants' ability to realize scalar inferences online. That prediction was borne out: whereas in Experiment 4 participants were able to realize scalar inferences and use that information to facilitate integration of "the rest" later in the sentence, in this experiment the facilitation effect disappeared. This suggests that the realization of scalar inferences is sensitive to the availability of processing resources. This complements the findings of previous research that have shown overt, strategic judgments of underinformative sentences to be sensitive to task demands (De Neys & Shaeken, 2007; Dieussaert et al., 2011) and time pressure (Chevallier et al., 2008; Bott et al., 2012), and extends those findings by further suggesting that it is not just strategic evaluation, but also implicit processes related to inferencing, that are affected by processing load.

3.5.3.1. The relationship between the secondary task and inferencing

The precise nature of the influence that the secondary task has on inferencing, however, is unclear. The possibility suggested above is that in the presence of a concurrent processing load, participants were less able to *realize* scalar inferences. There are at least three alternative explanations, however: 1) participants were less able to recognize the information-structural constraints of the different contexts; 2) participants were less able to *use* the upper-bounded interpretation of *some of* to facilitate access and/or integration of "the rest"; and 3) participants realized the inference and were then unable to cancel it in the face of extra processing load.

Regarding the first possibility, it should be noted that in the present experiment, the secondary task (listening to irrelevant speech) was ongoing throughout the entirety of the reading task, including when participants were reading the sentences that established the context as being upper- or lower-bounded. Thus, it is possible that scalar inferencing itself was not affected by the presence of the concurrent task, but that participants under the load conditions were simply less able to recognize the boundedness of the contexts, and thus unable to use contextual information to decide whether to inference or not to inference. This alternative seems unlikely, given that several effects of Boundedness were observed in the reading times, just no effects that were related to inferencing in particular. Nevertheless, this remains an alternative hypothesis that must be ruled out empirically in the future.

The second alternative explanation is that participants did indeed realize inferences in the upper-bounded contexts even under concurrent processing load, but the load prevented them from using the inference-based information in such a way that would facilitate reading times at "the rest" downstream. It is unclear whether concurrent processing load should interfere with basic lexical access and integration, but there is empirical evidence that concurrent processing load influences individuals' ability to comprehend sentences with increasing numbers of propositions, for example (see Caplan & Waters, 1999). Therefore, it is not possible to rule this possibility out on the basis of the present data.

The third possibility is that the equivalence of reading times for "the rest" in both context was not due to participants' failing to realize the inference in either context, but due to their realizing the inference in both contexts and failing to cancel it in the lower-bounded context. Such a finding would be consistent with default accounts of scalar inferencing, if such accounts assume that the cancellation of a scalar inference is an effortful process (see further discussion of this point in Section 3.6, General Discussion). A potential piece of evidence for this interpretation comes from the comparison between reading times in the *some of* and *only some of* sentences. One might predict that if the concurrent processing load only influenced the ability to *realize* inferences, then reading times for "the rest" would be slower in both contexts following *some of* (since the inference was not realized in either context) and faster in both contexts following *only some of* (since no pragmatic inference is required to realize the upper-bounded interpretation of this phrase; Minai & Fiorentino, 2010). This was not the case; in fact, reading times for "the rest" were numerically *faster* following *some of* than *only some of* (b = -0.018, SE = 0.01, t = -1.54, p = .136), which is more consistent with the notion that the upper-bounded meaning was indeed realized. It should be noted, however, that the comparison between segments of *some of* sentences and segments of the *only some of* sentences is not straightforward; although I have presented inference realization and subsequent facilitation of "the rest" as an all-or-nothing phenomenon, it is also possible that the strength of the upper-bounded interpretation based on scalar inference differs from that based on semantic composition with *only*, meaning that strong conclusions should not be drawn based on a direct comparison such as this.

In short, it is difficult to conclude on the basis of this experiment whether the presence of a concurrent processing load interfered with inferencing itself, or interfered with other processes related to the upper-bounded meaning. Nevertheless, the results of this experiment do indicate that the addition of a processing load modulates the context-sensitivity of scalar inferencing, and that previous findings regarding the influence of processing load on interpretation of underinformative sentences may not be just due to offline verification processes.

3.5.3.2. Comparing different background speech conditions

It is also worth noting that, contrary to expectations, the different background speech conditions did not elicit qualitatively different results in terms of their influence on scalar inferencing. It is likely that both conditions were difficult enough for participants that any differences between their effects on scalar inferencing were masked by a floor effect. Recall that the original prediction, based on Martin et al. (1988), was that real-word speech would create a larger processing load than novel-word speech. The present study differed from that study in at least three respects. First, real-word and novel-word background trials were randomly intermixed in the present study, whereas Martin and colleagues (1988) used a block design. The present study used a randomized design in order to be able to make stronger claims about how different background speech conditions might affect inferencing online (in a blocked or between-participants design, a difference between background conditions might be due to differences in conscious strategies people adopt in different blocks); this, however, could have reduced potential differences between the conditions, if listeners normally need several trials to retune their processing system for a particular kind of background speech, and the randomization could have introduced extra processing costs if participants needed to spend part of every trial determining whether the background speech was real or novel. The second difference is that the present study used multi-talker recordings for background speech, whereas the previous study used single-talker recordings; this was done in order to avoid introducing a potential influence of the list-like rhythm of single-talker stimuli on participants' self-paced reading pace, but this manipulation may have introduced differences between the present study and the previous study. The third difference is that the present study used self-paced reading, whereas participants in the previous study read sentences naturally, presented in full; self-paced reading may involve an additional processing component which is not present in natural reading and which may interact differently with the background speech conditions.

It is unclear which condition was actually more difficult in the present study, given that novel-word speech caused higher error rates and real-word speech caused higher reading times (at least in lower-bounded contexts). This could be an instance of a speed-accuracy trade-off, if participants were hurrying to quickly read past the "difficult" real-word conditions. The reading time results are difficult to interpret, however; Martin and colleagues' (1988) finding that real-word speech was more difficult was based on that condition's influence on comprehension accuracy, rather than reading times, and it is not straightforward to predict whether "more difficult" background speech would lead to slower or to faster reading times. Qualitatively, several participants reported that they found the novel-word speech more distracting, and several other participants reported that they found the real-word speech more distracting.

3.5.3.3. The speed of inferencing under concurrent processing load

Another result from the present experiment that bears mention is the interaction between lag time and the context effect at "the rest". Recall that Experiment 4 did not replicate the lag time effect that Hartshorne and Snedeker (submitted) report: whereas that study found that inferences only facilitated reading times at "the rest" in a long-lag condition, Experiment 4 of the present dissertation showed no moderation of the facilitation effect by lag time. In the present experiment, however, an effect similar to that described by Hartshorne & Snedeker (submitted) was observed in the novel-word background condition: no facilitation expect was observed at short lag times, but at long lag times the facilitation effect (shorter reading times for "the rest" in upper-bounded rather than lower-bounded contexts) did emerge. This could be interpreted as evidence for delayed inferencing in this condition. On the other hand, no facilitation effect emerged at any lag time in the real-word background speech condition. Comparing these results to the result from Experiment 4, one might speculate that inferencing could occur rapidly when the parser is relatively unburdened (no background speech), at a delay when the parser is somewhat burdened (novel-word background speech), and not at all when the parser is even more burdened (real-word background speech). Note, however, that this interpretation requires the assumption that real-word background speech was more taxing for the parser than novel-word background speech; that assumption is supported by the findings of Martin and colleagues (1988) and the overall reading time data of the present experiment, but not by the accuracy data of the present experiment. As mentioned above, which background speech condition was more difficult remains an open question.

3.5.3.4. Individual differences in inferencing under concurrent processing load

Finally, the present experiment identified several individual-level cognitive factors that may be of relevance to scalar inferencing. At the quantifier, there was a trend towards participants with higher working memory (as measured by the Counting Span task) showing longer reading times in the upper-bounded than the lower-bounded context. This is the effect predicted by context-driven accounts of inferencing—i.e., that realizing an inference will elicit a processing effort—that was not observed in the previous experiment. One possible interpretation of this pattern of effects is that inferencing occurred immediately and effortlessly in the previous experiment, where there was no concurrent speech, whereas in the present experiment the concurrent task made participants with low working memory unable to realize the inferences rapidly, and participants with high working memory only able to realize inferences with substantial effort. Some aspects of the data, however, are inconsistent with this account. Firstly, working memory did not modulate the context effect at any other points in the sentence; if participants with different amounts of working memory resources differed in their ability to realize inferences, then one would expect to observe more, later effects of working memory (either as the participants with low working memory caught up with the others and realized the inference later, or at "the rest" where high-WM participants might show a facilitation due to inference whereas low-WM participants might not). Secondly, in a study by Dieussaert and colleauges (2011), it was participants with *low* working memory (as measured by the Operation Span, another working memory test that is not based on reading) whose inferencing was modulated by concurrent processing load. It is unclear why Counting Span would show a stronger relationship with inferencing and ignoring background speech but Reading Span would not; the processing components of these tasks involve different cognitive demands, however, so the fact that this effect was not observed in the Reading Span could potentially be informative about the nature of the cognitive demands imposed by inferencing in certain contexts. The proposal described above also necessitates the assumption that the effect of concurrent background speech is not incurred directly on the portion of working memory measured by the Counting Span. If it were, that would mean that having less working memory resources is similar to having a concurrent background task—in which case one would predict low-span participants in Experiment 4 to show a slowdown at this segment like high-span partcipants in the current experiment did.

The other individual difference measure of interest was the Perspective-Taking subscale of the Interpersonal Activity Index. Recall that participants in Experiment 4 showed a facilitation effect at "the rest" (faster reading times in the upper-bounded context, where the scalar inference had been realized, than in the lower-bounded context), but participants in the present experiment did not show such a facilitation effect in the group analysis, which suggested that the presence of a concurrent processing load interfered with either their ability to realize inferences or their ability to cancel inferences (see above for other possible accounts). In fact, however, participants with a high perspective-taking ability did show a trend towards having this facilitation effect. This suggests that perspective-taking ability is related to the ability to perform inference-related processing in cognitively taxing situations. It remains difficult to tell, however, whether the task modulated participants' abilities to realize inferences or to cancel them. In an fMRI study by Sai and colleagues (submitted), perspective-taking ability on the IRI was correlated with BOLD activations in the bilateral inferior frontal gyrus possibly related to inhibiting inferences. They compared *even* sentences that are congruent with the inference made from noun modified by *even* (e.g., "Even such a quiet sound, he can hear", in which the *even*-NP triggers an inference that the sentence will be about someone's good hearing ability) to *even* sentences that are incongruent with the inference (e.g., "Even such a loud sound, he can hear", in which the *even*-NP triggers an inference that the sentence will be about someone's *poor* hearing ability), and found that participants with higher perspective-taking ability showed the highest activation in the latter condition, compared to the former. They concluded that the incongruent sentences required inhibition of an inference, and that inhibition was modulated by perspective-taking ability. (It is an open question, however, whether participants with higher perspective-taking ability showed more activation because they performed this inhibition more, or because it took them more effort to perform it.) Perspective-taking ability is also correlated with gray matter volume in the anterior cingulate cortex (Banissy et al., 2012), a region involved in conflict monitoring and social cognition which has also been implicated in the processing of incongruous inferences (Shetreet et al., in press). Under such an account, the relationship between perspective-taking and the context effect in the present experiment might be taken as evidence that the inference was realized rapidly and by default in all contexts, and only participants with high perspective-taking ability were then able to cancel it in lower-bounded contexts (thus leading to the facilitation effect for these participants), whereas participants with low perspective-taking ability were unable to do so. On the other hand, whereas Sai and colleagues (submitted) implicate the bilateral inferior frontal gyrus (the region in which they observed correlations with perspective-taking ability) in inference cancellation, Shetreet and colleagues (in press) implicate the left inferior frontal gyrus in inference *realization*. If perspective-taking abilities play a role in inference realization rather than inference cancellation, the same pattern of results in the present study could be taken as evidence that the inference was realized by all participants in Experiment 4, but that in the present experiment it was realized only by participants with high perspective-taking ability, and was not realized at all by participants with low perspective-taking ability. In short, further investigation is needed to elucidate the relationship between perspective-taking ability and the realization of scalar inferences.

3.6. GENERAL DISCUSSION

The experiments reported in this chapter yielded three main results. First, the realization of scalar inferences was sensitive to the information-structural constraints of the context—such that inferences were realized when the meaning they contribute would be discursively relevant, and not realized when it would not be. Second, realizing scalar inferences did not elicit directly observable processing costs in omnibus analyses. Third, inferencing was sensitive to the availability of processing resources, such that the context-sensitivity of inferencing was not observed when participants were under a concurrent processing load.

The present results raise questions for context-based models. While numerous recent studies have suggested that inferences are realized at a delay except in special contexts (e.g., Huang & Snedeker, 2009; Bott et al., 2012; Hartshorne and Snedeker, submitted), the traditional explanation for that finding is that inferencing is effortful and thus the parser avoids inferencing until after it can evaluate whether the extra effort is worthwhile, or at least until after the core semantic meaning of the scalar term has already been realized. The context-driven accounts' predictions about the delayed realization of scalar inferences are still tenable without evidence for processing costs—such accounts assume that the output of semantic composition feeds into the inferencing process, and thus even if inferencing itself is effortless it cannot be done until after semantic composition is complete—but it is unclear how such accounts could explain the context-sensitivity without recourse to processing costs. If inferencing is not effortful, then a new explanation for the context-sensitivity would be needed (see Bott et al., 2012, for several alternative accounts). Alternatively, inferencing may be effortful but reading times may not be sensitive to this effort. If that is the case, future studies must use other methods, such as event-related potentials, to test for different instantiations of processing costs.

The present study also raises questions for default accounts—specifically, while a default model could account for the present findings (by assuming that the inference was effortlessly realized at "some of" and then cancelled in the lower-bound context before "the rest"), default models owe an account of the nature of inference cancellation and the processes that underlie it. Levinson (2000: 49-54) describes two algorithms for determining whether a default inference will be cancelled. The first involves checking whether an inference is consistent with the previous context or higher-ranked information (e.g., in the statement "some of the students are hardworking; in fact, all of them are", the inference "not all of the students are hardworking" is inconsistent with the explicit entailment "all of the students are [hardworking]"—in this formulation, entailments take precedence over implicatures).[33] The lower-bounded contexts in

[33] Katsos and Cummins (2010: 286) make reference to additional epistemic factors in the context which could cause an inference to be cancelled or not realized, such as if a speaker is known to be non-cooperative.

the present study would not trigger inference cancellation from this mechanism, since the inference does not conflict with information in the sentence or prevent the comprehender from completing the task (i.e., the question of "whether *any* of John's relatives are staying in his apartment" is answered even if the answer is "*some but not all* of them are"). Therefore, the fact that the inference was cancelled in lower-bound contexts before "the rest" (as evidenced by slower reading times to "the rest" in that context in Experiment 4) would have to be explained through the second cancellation mechanism described by Levinson (2000), whereby inferences that are irrelevant to the goal of the conversation are discarded. However, Breheny, Katsos, and Williams (2006; see also Katsos & Cummins, 2010: 287, 288) assume that inference cancellation should involve extra effort, and some experimental evidence also suggests that it does (Feeney et al., 2004; Politzer-Ahles et al., 2013). If the processor avoids unnecessary effort, it is unclear why it would make the effort to cancel inferences that do not interfere with the comprehension of the utterance. As suggested by Levinson (2000: 53), the default model is lacking a full account of what about this particular context would cause inference cancellation, and the nature of the process through which inferences are cancelled; the results of the present study highlight the need for such an account if the default model is to explain how meaning is realized online in the contexts tested in this experiment.

The present results may be amenable to the constraint-based account proposed by Degen and Tanenhaus (2011). Under this account, scalar inferencing is a result of rapid integration of multiple constraints, which may facilitate or inhibit the inference. Unlike traditional context-driven accounts, this account may predict that inferencing is both context-sensitive and potentially rapid and effortless. If numerous constraints strongly facilitate the inference, then realizing the inference may not require great effort; on the other hand, if constraints discourage the comprehender from making the inference, it may not be realized at all. Such a model would be able to account for seemingly effortless inferencing in contexts like the upper-bound context of the present study. This is different from traditional context-driven models, which assume that inferencing is always costly and therefore that when it does happen it will be late and effortful. Further study would be useful to investigate the predictions of a constraint-based account for this type of paradigm.

A constraint-based approach may also offer an explanation for the puzzling difference between the results of the present experiments and those of the experiment reported by Bergen and Grodner (2012). Recall that in that experiment, a reading time slowdown *was* observed at the quantifier in the context that supports scalar inferencing, and it was not due to the sorts of lexical confounds that were present in the study by Breheny and colleagues (2006). At face value, the results of that experiment seem to contract this one. That experiment, however, used a different context manipulation than the present experiment. Experiments using information structure (upper- vs. lower-bounding) as a context manipulation have not robustly found evidence for directly observable processing costs from inferencing (Lewis & Phillips, 2011; this study), nor have experiments using semantic structure (entailment polarity) as the manipulation (Hartshorne & Snedeker, submitted), whereas Bergen and Grodner's (2012) study using knowledge of the speaker's epistemic state has found such evidence. It is possible that these context manipulations facilitate inferencing to greater or lesser extents. In the constraint-based framework, a context that facilitates inferencing only to a small extent may lead to a case in which inferencing occurs but it is not entirely immediate or effortless.

A remaining question concerns the significance of the fact that the context-sensitivity of inferencing was eliminated under processing load. As described in the previous section, this result may be consistent with any of the processing models described here, depending on whether the processing load prevented inferencing itself, or inference cancellation, or the use of inference-based information to assist in lexical access and discourse integration. Further study is necessary to elucidate the role of concurrent processing load in the realization of scalar inferences.

In conclusion, the experiments presented in this chapter raise questions for both traditional accounts of inferencing, and suggests that alternative accounts or reformulations of these accounts may be worth considering. The results also challenge the field to seek evidence for processing costs in new ways. Both of these endeavors have the potential to improve our understanding of how comprehenders compose the meaning of utterances in real-time.

CHAPTER 4: CONCLUSIONS

The experiments reported here examined the representation processing of inference-based meaning using neurolinguistic and psycholinguistic techniques. The first set of experiments (Experiments 1-3) was designed to test whether the processing of enriched, scalar inference-based meanings is subserved by cognitive and neural mechanisms that are independent from those subserving the processing of lexical and compositional semantic meaning. Previous electrophysiological studies of this type of meaning had used a paradigm which only allowed for the examination of how the scalar inference influenced the processing of later words, and thus did not provide opportunities to examine how the scalar inference itself was processed. Using a new picture-sentence verification design, the present experiments showed evidence that the processing of the scalar inference elicits a unique electrophysiological response at the position of the scalar term itself: unlike lexico-semantic and compositional semantic violations, scalar implicature-based violations elicited a broad sustained negativity. This ERP pattern may be related to revision and reinterpretation of meaning, which was possible in the scalar implicature-based violations but not the other violations. These experiments also expanded the empirical domain of scalar implicature research by testing Chinese, a language which has previously been the focus of only one experimental investigation of scalar implicature (Wu & Tan, 2009), whereas the rest of the research on online scalar implicature processing has been conducted almost entirely on Indo-European languages.

While those experiments examined how comprehenders accommodate a meaning that is incompatible with the context, the second series of experiments (Experiments 4-5) investigated how comprehenders realize that meaning in the first place. These experiments compared reading times to scalar quantifiers that received an upper-bounded interpretation to those that received a lower-bounded interpretation. The results showed that realizing the upper-bounded interpretation did not elicit any directly observable processing costs. This finding raised the question of why certain commonly-reported aspects of scalar inferencing, such as delay and context-sensitivity, would emerge if inferencing is not effortful. The results of these experiments challenged traditional context-driven accounts, which suppose properties such as delay and context-sensitivity are a direct result of the effort required to realize inferences. They may be consistent, however, with a constraint-based account that views inferencing as more gradient, and thus may be able to account for the possibility that inferencing seems to be effortful in some contexts and effortless in others.

The present experiments have also raised new research questions to be addressed. The ERP experiments reported in Chapter 2 provided evidence for the existence of an ERP correlate of pragmatic revision, but the precise neural substrates of this effect are still unknown. Identifying the neural generators of this effect may further our understanding of the processes underlying the revision or inhibition of inference-based meaning. In particular, a deeper understanding of the neural and cognitive mechanisms subserving scalar inference processing may shed light on the nature of scalar inferencing itself, and the controversial question of whether it is a "pragmatic" or "semantic" phenomenon (see the discussion in Section 1.1).

It also remains to be seen whether this effect is unique to the kind of inferences examined in this dissertation, or whether it generalizes to other pragmatic phenomena—if the hypothesis that it reflects revision or inhibition of a particular aspect of meaning is correct, then this ERP may also appear for similar phenomena, such as standardization implicitures (Garrett & Harnish, 2007).

The results of the self-paced reading experiments, along with those of previous self-paced reading and eye-tracking experiments, have raised the possibility that scalar inferencing may evoke processing costs that are not detectable through self-paced reading times. Thus, future work testing these kinds of phenomena for processing costs using other methods, such as ERPs, will be valuable for the field of experimental pragmatics. Another very interesting possibility highlighted by recent self-paced reading studies (particularly the comparison between the studies reported by Hartshorne & Snedeker, submitted; Bergen & Grodner, 2012; and here) is that different contexts differ in the strength of the bias they create for or against inferencing. This possibility underscores the need for the field to consider new measures that are sensitive to the strength of contextual biases and to the strength of the activations of each interpretation of *some of*, rather than all-or-nothing measures of whether the inference was made or not. Eye-tracking (visual world and look-and-listen) may be one such promising method, and has already been shown to be an effective tool for investigating scalar inferencing in numerous studies (Huang & Snedeker, 2009; Grodner et al., 2010; Panizza et al., 2011; Breheny et al., 2012, in press). However, visual world and look-and-listen data need to be aggregated over multiple trials to reveal gradient biases, and thus may not be sensitive to differences in bias across individual items (although they are sensitive to differences in bias between groups of items). Neurolinguistic methods may also be useful if the neural correlates of considering an upper-bounded meaning can be identified.

The present experiments focused exclusively on scalar inferences associated with quantifiers such as *some of*. For reasons described in the first chapter, there are sound methodological reasons for choosing this scalar, which is why it has become such a popular case for investigation in experimental pragmatics: the different interpretations of *some of* map onto quantifiable alternatives (e.g., *all of* or *not all of*) which are easy to map onto discrete representations in the real world, which makes it relatively easy to construct distinct conditions for experimental analysis. Nonetheless, other linguistic expressions, such as *or*, have scalar inferences associated with them. Furthermore, scalar inferences can be associated with nearly any linguistic expression if the context allows the comprehender to form a post-hoc scale. For instance, in a context in which there are several forks, several spoons, and a box, the utterance "The woman put a fork into the box" may be understood as meaning "The woman put a fork *and nothing else* into the box", where the *and nothing else* interpretation is derived through a scalar implicature specific to that context. These inferences that are based on a context rather than a specific linguistic expression like *some of* are often referred to as *particularized conversational implicatures* (although accounts based on Relevance Theory assume that all implicatures, including those about expressions like *some of*, are cases of particularized conversational implicature; see Noveck & Sperber, 2007; Katsos & Cummins, 2010). It is indeed possible to conduct well-controlled investigation of scalar expressions such as *or* (see, e.g., Breheny et al., 2006; Chevallier et al., 2008), and of particularized conversational implicatures (Breheny et al., in press). Given that scalar implicature does not seem to be a monolithic phenomenon, but rather that different scales seem to differ greatly in the strength of the upper-bounded meanings they create and the fuzziness of elements on the scale (Doran et al., 2009, 2012; Newstead, 1988), much work remains to test whether conclusions made about the speed, effort, and context-dependency of inferencing related to *some of* will extend to other types of inferencing or not.

In summary, the present dissertation has presented several experimental approaches to investigating the processing of scalar inferencing, and has shown that the mechanisms subserving scalar implicature processing may be more gradient and constraint-based than traditional accounts assumed.While a wide variety of contexts and manipulations remains to be tested before the field can arrive at a full understanding of the general mechanisms subserving scalar inferencing, the neurolinguistic and psycholinguistic approach described in this dissertation presents a way forward in studying how this important aspect of meaning is realized and negotiated during the real-time comprehension of natural language.

REFERENCES

Baayen, H., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390-412.

Baggio, G., van Lambalgen, M., & Hagoort, P. (2008). Computing and recomputing discourse models: An ERP study. *Journal of Memory and Language, 59,* 36-53.

Banissy, M., Kanai, R., Walsh, V., & Rees, G. (2012). Inter-individual differences in empathy are reflected in human brain structure. *NeuroImage, 62*, 2034-2039.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): evidence from Asperger Syndrome/high-functioning Autism, males and females, scientists and mathematicians. *Journal of Autism ad Developmental Disorders, 31*, 5-18.

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy, 4*, 159-219.

Bergen, L., & Grodner, D. (2012). Speaker knowledge influences the comprehension of pragmatic inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 1450-1460.

Bezuidenhout, A., & Cutting, J. (2002). Literal meaning, minimal propositions and pragmatic processing. *Journal of Pragmatics, 34*, 433-456.

Boersma, P., & Weenink, D. (2012). Praat: doing phonetics by computer [Computer program]. http://www.praat.org/

Bonin, P., Peereman, R., Malardier, N., Méot, A., & Chalard, M. (2003). A new set of 299 pictures for psycholinguistic studies: French norms for name agreement, image agreement, conceptual familiarity, visual complexity, image variability, age of acquisition, and naming latencies. *Behavior Research Methods, Instruments, & Computers, 35,* 158-167.

Bornkessel-Schlesewsky, I., Kretzschmar, F., Tune, S., Wang, L., Genç, S., Philipp, M., … Schlesewsky, M. (2011). Think globally: Cross-linguistic variation in electrophysiological activity during sentence comprehension. *Brain and Language, 117,* 133-152.

Bott, L., Bailey, T., & Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. *Journal of Memory and Language, 66*, 123-142.

Bott, L., & Noveck, I. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language, 51,* 437-457.

Breheny, R., Ferguson, H., & Katsos, N. (2013). Taking the epistemic step: toward a model of on-line access to conversational implicatures. *Cognition, 126*, 423-440.

Breheny, R., Ferguson, H., & Katsos, N. (in press). Investigating the timecourse of accessing conversational implicatures during incremental sentence interpretation. *Language and Cognitive Processes*.

Breheny, R., Katsos, N., & Williams, J. (2006). Are generalized scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition, 100,* 434-463.

Calvo, M. (2001). Working memory and inferences: evidence from eye fixations during reading. *Memory, 9*, 365-381.

Caplan, D., & Waters, G. (1999). Verbal working memory and sentence comprehension. *Behavioral and Brain Sciences, 22*, 77-94.

Chemla, E., & Spector, B. (2011). Experimental evidence for embedded scalar implicatures. *Journal of Semantics, 28*, 359-400.

Chevallier, C., Noveck, I., Nazir, T., Bott, L., Lanzetti, V., & Sperber, D. (2008). Making disjunctions exclusive. *Quarterly Journal of Experimental Psychology, 61*, 1741-1760.

Chi, W. (2000). "Bùfen", "yǒu de" zhī luójì biànxī. [Towards a logical differentiation between "part" and "some"]. *Shāndōng Shīdà Xuébào (Shèhuì Kēxué Bǎn) [Shandong Normal University Journal (Social Science)], 169,* 91-103.

Chierchia, G. (2004). Scalar implicatures, polarity phenomena and the syntax/pragmatics interface. In A. Belletti (Ed.), *Structures and Beyond* (pp. 39-103). Oxford: Oxford University Press.

Chierchia, G., Fox, D., & Spector, B. (2012). Scalar implicature as a grammatical phenomenon. In C. Maienborn, K. von Heusinger, & P. Portner (Eds.), *Semantics: An International Handbook of Natural Language Meaning* (pp. 2297-2332). New York: Mouton de Gruyter.

Conway, A., Kane, M., Bunting, M., Hambrick, D., Wilhelm, O., & Engle, R. (2005). Working memory span tasks: a methodological review and user's guide. *Psychonomic Bulletin and Review, 12*, 769-786.

Davis, M. (1983). Measuring individual differences in empathy: evidence for a multidimensional approach. *Journal of Personality and Social Psychology, 44*, 113-126.

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

Degen, J. (2009). *Processing scalar implicatures: An eye-tracking study* (Master's thesis, University of Osnabrück, Osnabrück, Germany).

Degen, J., & Tanenhaus, M. (submitted). Processing scalar implicature: a constraint-based approach.

Degen, J., & Tanenhaus, M. (2011). Making inferences; the case of scalar implicature processing. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 3299-3304).

Degen, J., & Tanenhaus, M. (2010). *When contrast is salient, pragmatic "some" precedes logical "some".* Poster presented at the 23rd CUNY Conference on Human Sentence Processing, New York, NY.

Dieussaert, K., Verkerk, S., Gillard, E., & Schaeken, W. (2011). Some effort for some: further evidence that scalar implicatures are effortful. *Quarterly Journal of Experimental Psychology, 64*, 2352-2367.

Doran, R., Baker, R., McNabb, Y., Larson, M., & Ward, G. (2009). On the non-unified nature of scalar implicature: an empirical investigation. *International Review of Pragmatics, 1*, 211-248.

Doran, R., Ward, G., Larson, M., McNabb, Y., & Baker, R. (2012). A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language, 88*, 124-154.

Eriksen, B., & Eriksen, C. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception and Psychophysics, 16*, 143-149.

Engle, R., Tuholski, S., Laughlin, J., & Conway, A. (1999). Working memory, short-term memory and general fluid intelligence: a latent variable approach. *Journal of Experimental Psychology: General, 128*, 309-331.

Feeney, A., Scafton, S., Duckworth, A., & Handley, S. J. (2004). The story of some: Everyday pragmatic inferences by children and adults. *Canadian Journal of Experimental Psychology, 54*, 128-133.

Filik, R, & Leuthold, H. (2008). Processing local pragmatic anomalies in fictional contexts: evidence from the N400. *Psychophysiology, 45*, 554-558.

Fischler, I., Bloom, P., Childers, D., Roucos, S., & Perry, N. (1983). Brain potentials related to stages of sentence verification. *Psychophysiology, 20*, 400-409.

Fiorentino, R., Politzer-Ahles, S., & Pak, N. (2012). Probing the dynamics of complex word recognition: An ERP investigation of the processing of novel compounds. *Poster presented at the 4th Neurobiology of Language Conference*. San Sebastian, Spain.

Fiorentino, R., Politzer-Ahles, S., Popescu, E., & Popescu, M. (2011). The role of morphological juncture identification in complex word processing: an MEG investigation. *Poster presented at the 41st Society for Neuroscience Annual Meeting*. Washington, DC.

Foppolo, F. (2007). Between 'cost' and 'default': a new approach to scalar implicature. *Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue,* 125-132.

Gadzar, G. (1979). *Pragmatics: Implicature, presupposition and logical form*. New York: Academic Press.

Garrett, M, & Harnish, R. (2007). Experimental pragmatics: testing for implicitures. *Pragmatics and Cognition, 15*, 65-90.

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

Geurts, B., & van Tiel, B. (to appear). Embedded scalars. *Semantics and Pragmatics.*

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41-58). New York: Academic Press.

Grodner, D., Gibson, E., & Tunstall, S. (2002). Syntactic complexity in ambiguity resolution. *Journal of Memory and Language, 46*, 267-295.

Grodner, D., Klein, N., Carbary, K., & Tanenhaus, M. (2010). "Some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition, 116,* 42-55.

Groppe, D., Makeig, S., & Kutas, M. (2009). Identifying reliable independent components via split-half comparisons. *NeuroImage, 45*, 1199-1211.

Hackl, M. (2009). On the grammar and processing of proportional quantifiers: *most* versus *more than half. Natural Language Semantics, 17*, 63-98.

Hagoort, P. (2003). Interplay between syntax and semantics during sentence comprehension: ERP effects of combining syntactic and semantic violations. *Journal of Cognitive Neuroscience, 16,* 883-899.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. (2004). Integration of word meaning and world knowledge in language comprehension. *Science, 304*, 438-441.

Hagoort, P., & Van Berkum, J. (2007). Beyond the sentence given. *Philosophical Transactions of the Royal Society B, 362*, 801-811.

Hartshorne, J., & Snedeker, J. (submitted). The speed of inference: Evidence against rapid use of context in calculation of scalar implicatures.

Horn, L. (1972). *On the semantic properties of the logical operators in English*. Ph.D. thesis, UCLA.

Horn, L. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In D. Schiffrin (Ed.), *Meaning, form, and use in context: Linguistic applications* (pp. 11-42). Washington: Georgetown University Press.

Huang, Y., Hahn, N., & Snedeker, J. (2010). *Some inferences still take time: Prosody, predictability, and the speed of scalar implicatures.* Poster presented at the 23rd CUNY Conference on Human Sentence Processing, New York, NY.

Huang, Y., & Snedeker, J. (2011). Language and conversation revisited: evidence for a division between semantic and pragmatic content in real-time language comprehension. *Language and Cognitive Processes, 26*, 1161-1172.

Huang, Y., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: insight into the semantics-pragmatics interface. *Cognitive Psychology, 58*, 376-415.

Hunt, L., Politzer-Ahles, S., Gibson, L., Minai, U., & Fiorentino, R. (2013). Pragmatic inferences modulate N400 during sentence comprehension: Evidence from picture-sentence verification. *Neuroscience Letters, 534*, 246-251.

Ippolito, M. (2011). A note on embedded implicatures and counterfactual presuppositions. *Journal of Semantics, 28*, 267-278.

Jenkins, A., & Mitchell, J. (2010). Mentalizing under uncertainty: dissociated neural responses to ambiguous and unambiguous mental state inferences. *Cerebral Cortex, 20*, 404-410.

Jiang, X., Tan, Y., & Zhou, X. (2009). Processing the universal quantifier during sentence comprehension: ERP evidence. *Neuropsychologia, 47*, 1799-1815.

Jing, H., Pivik, R., & Dykman, R. (2006). A new scaling method for topographical comparisons of event-related potentials. *Journal of Neuroscience Methods,* 151, 239-249.

Just, M., Carpenter, P., & Woolley, J. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General, 111*, 228-238.

Kane, M., Hambrick, D., Tuholski, S., Wilhelm, O., Payne, T., & Engle, R. (2004). The generality of working memory capacity: a latent-variable approach to verbal and visuo-spatial memory span and reasoning. *Journal of Experimental Psychology: General, 133*, 189-217.

Katsos, N., & Cummins, C. (2010). Pragmatics: from theory to experiment and back again. *Language and Linguistics Compass, 4,* 282-295.

Katsos, N., Roqueta, C., Estevan, R., & Cummins, C. (2011). Are children with Specific Language Impairment competent with the pragmatics and logic of quantification? *Cognition, 119*, 43-57.

Knoeferle, P., Urbach, T., & Kutas, M. (2011). Comprehending how visual context influences incremental sentence processing: insights from ERPs and picture-sentence verification. *Psychophysiology, 48*, 495-506.

Kuperberg, G., McGuire, P., Bullmore, E., Brammer, M., Rabe-Hesketh, S., Wright, I., … David, A. (2000). Common and distinct neural substrates for pragmatic, semantic, and syntactic processing of spoken sentences: an fMRI study. *Journal of Cognitive Neuroscience, 12*, 321-341.

Kutas, M., & Federmeier, K. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Science, 4,* 463-470.

Kutas, M., Van Petten, C., & Kluender, R. (2006). Psycholinguistics electrified II: 1994-2005. In M. Traxler & M.A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (2nd ed., pp. 659-724). New York: Elsevier.

Lau, E., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (De)constructing the N400. *Nature Reviews Neuroscience, 9,* 920-933.

Leuthold, H., Filik, R., Murphy, K., & Mackenzie, I. (2012). The on-line processing of socio-emotional information in prototypical scenarios: inferences from brain potentials. *Social Cognitive and Affective Neuroscience, 7*, 457-466.

Lewis, S., & Phillips, C. (2011). Computing scalar implicatures is cost-free in supportive contexts. *Poster presented at 17th Annual Conference on Architectures and Mechanisms for Language Processing*.

Levinson, S. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge: MIT Press.

Li, C., & Thompson, S. A. (1981). *Mandarin Chinese: A Functional Reference Grammar*. Los Angeles: University of California Press.

Li, T. (1983). Distribution of left/right handedness among Chinese people. *Xin Li Xue Bao [Acta Psychologica Sinica], 15*, 268-276.

Lu, A., & Zhang, J. (2012). Event-related potential evidence for the early activation of literal meaning during comprehension of conventional lexical metaphors. *Neuropsychologia,* 50, 1730-1738.

Makeig, S., Bell, A., Jung, T., & Sejnowski, T. (1996). Independent component analysis of electroencephalographic data. In D. Touretzky, M. Mozer, & M. Hasselmo (Eds.), *Advances in neural information processing systems 8* (pp. 145-151). Cambridge: MIT Press.

Martin, R., Wogalter, M., & Forlano, J. (1988). Reading comprehension in the presence of unattended speech and music. *Journal of Memory and Language, 27*, 382-398.

Meibauer, J. (2012). Pragmatic evidence, context, and story design: an essay on recent developments in experimental pragmatics. *Language Sciences, 34*, 768-776.

Minai, U., & Fiorentino, R. (2010). The role of the focus operator *only* in children's computation of sentence meaning. *Language Acquisition, 17*, 183-190.

Morgan, B., Gross, G., Clark, R., Dreyfuss, M., Boller, A., Camp, E., … Grossman, M. (2011). Some is not enough: quantifier comprehension in corticobasal syndrome and behavioral variant frontotemporal dementia. *Neuropsychologia, 49*, 3532-3541.

Newstead, S. (1988). Quantifiers as fuzzy concepts. In T. Zétényi (Ed.). *Advances in Psychology, 56* (pp. 51-72). New York: North-Holland.

Nicolle, S. (2003). Mental models theory and relevance theory in quantificational reasoning. *Pragmatics and Cognition, 11*, 345-378.

Nieuwland, M., Ditman, T., & Kuperberg, G. (2010). On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities. *Journal of Memory and Language, 63*, 324-346.

Nieuwland, M., & Kuperberg, G. (2008). When the truth is not too hard to handle: an event-related potential study on the pragmatics of negation. *Psychological Science, 19*, 1213-1218.

Nieuwland, M. & Van Berkum, J. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience, 18*, 1098-1111.

Noveck, I., & Posada, A. (2003). Characterizing the time course of an implicature: An evoked potentials study. *Brain Research, 85*, 203-210.

Noveck, I., & Sperber, D. (2007). The why and how of experimental pragmatics: The case of 'scalar inferences'. In N. Burton-Roberts (Ed.), *Advances in pragmatics* (pp. 184-212). Basingstoke: Palgrave.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia, 9*, 97-113.

Osterhout, L., Bersick, M. & McLaughlin, J. (1997). Brain potentials reflect violations of gender stereotypes. *Memory and Cognition, 25*, 273-285.

Panizza, D., Huang, Y., Chierchia, G., & Snedeker, J. (2011). Relevance of polarity for the online interpretation of scalar terms. *Proceedings of the 19th Semantics and Linguistic Theory Conference (April 2009)*, 360-378.

Pijnacker, J., Geurts, B., van Lambalgen, M., Buitelaar, J., & Hagoort, P. (2011). Reasoning with exceptions: an event-related brain potentials study. *Journal of Cognitive Neuroscience, 23*, 471-480.

Politzer-Ahles, S., & Fiorentino, R. (in press). The realization of scalar inferences: context sensitivity without processing cost. *PLoS ONE.*

Politzer-Ahles, S., & Fiorentino, R. (forthcoming). Sensitivity of online scalar inferencing to context and to processing load. *Poster to be presented at 5th Experimental Pragmatics Conference*.

Politzer-Ahles, S., Fiorentino, R., Jiang, X., & Zhou, X. (2013). Distinct neural correlates for pragmatic and semantic meaning processing: An event-related potential investigation of scalar implicature processing using picture-sentence verification. *Brain Research, 1490*, 134-152.

Politzer-Ahles, S., Jiang, X., Fiorentino, R., & Zhou, X. (2012). Individual differences in logical ability predict ERP responses to underinformative sentences. *Poster presented at 4th Neurobiology of Language Conference*. San Sebastian, Spain.

Pylkkänen, L., Brennan, J., Bemis, D. (2011). Grounding the cognitive neuroscience of semantics in linguistic theory. *Language and Cognitive Processes, 26*, 1317-1337.

Rullman, H., & You, A. (2006). General number and the semantics and pragmatics of indefinite bare nouns in Mandarin Chinese. In K. von Heusinger & K. P. Turner (Eds.), *Where semantics meets pragmatics* (pp. 175-196). Amsterdam: Elsevier.

Sai, L., Jiang, X., Yu, H., & Zhou, X. (submitted). Cognitive empathy modulates the processing of pragmatic constraints during sentence comprehension.

Shetreet, E., Chierchia, G., & Gaab, N. (in press). When *some* is not *every*: dissociating scalar implicature generation and mismatch. *Human Brain Mapping*.

Sperber, D., & Wilson, C. (1995). *Relevance: Communication and Cognition*. Oxford: Blackwell.

Steinhauer, K., Drury, J., Portner, P., Walenski, M., Ullman, M. T. (2010). Syntax, concepts, and logic in the temporal dynamics of language comprehension: Evidence from event-related potentials. *Neuropsychologia, 48*, 1525-1542.

Stroop, J. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643-662.

Szekely, A., Jacobsen, T., D'Amico, S., Devescovi, A., Andonova, E., Herron, D., … Bates, E. (2004). A new on-line resource for psycholinguistic studies. *Journal of Memory and Language, 51*, 247-250.

Tavano, E. (2010). *The balance of scalar implicature* (Doctoral dissertation, University of Southern California, Los Angeles, CA).

Tesink, C., Buitelaar, J., Petersson, K., van der Gaag, R., Kan, C., Tendolkar, I., & Hagoort, P. (2009). Neural correlates of pragmatic language comprehension in autism spectrum disorders. *Brain, 2009*, 1941-1952.

Trueswell, J., Tanenhaus, M., & Garnsey, S. (1994). Semantic influences on parsing: use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language, 33*, 285-318.

Tsai, W.-T. (2004). Tán "yǒu rén," "yǒu de rén," hé "yǒu xiē rén" [On "yǒu rén," "yǒu de rén," and "yǒu xiē rén"]. *Hànyǔ Xuébào [Chinese Linguistics], 8*(2), 16-25.

Urbach, T., & Kutas, M. (2010). Quantifiers qualify more or less online: ERP evidence for partial incremental interpretation. *Journal of Memory and Language, 63*, 158-179.

Van Berkum, J. (2009). The neuropragmatics of 'simple' utterance comprehension: An ERP review. In U. Sauerland & K. Yatsushiro (Eds.), *Semantics and Pragmatics: From Experiment to Theory* (pp. 276-316). Basingstoke: Palgrave Macmillan.

Van Berkum, J., Koornneef, A., Otten, M., & Nieuwland, M. (2007). Establishing reference in language comprehension: An electrophysiological perspective. *Brain Research, 1146*, 158-171.

Waters, G., & Caplan, D. (1996). The measurement of verbal working memory capacity and its relation to reading comprehension. *The Quarterly Journal of Experimental Psychology, 49A*, 51-79.

Wiswede, D., Koranyi, N., Müller, F., Langner, O., & Rothermund, K. (in press). Validating the truth of propositions: behavioral and ERP indicators of truth evaluation processes. *Social Cognitive and Affective Neuroscience.*

Wu, Z., & Tan, J. (2009). Hànyǔ értóng yǔyán zhōng de děngjí hányì—yí xiàng shíyàn yánjiū [Scalar implicature in Chinese child language: An experimental study]. *Wàiguóyǔ [Journal of Foreign Languages], 32*, 69-75.

Xie, Y. (2003). Guānyú "yǒu de+VP" [On the construction of "yǒu de+VP"]. *Yǔyán Yánjiū [Studies in Language and Linguistics], 23*, 37-4.

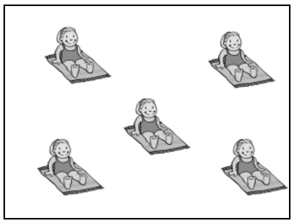
Ye, Z., & Zhou, X. (2009a). Conflict control during sentence comprehension: fMRI evidence. *NeuroImage, 48*, 280-290.

Ye, Z., & Zhou, X. (2009b). Executive control in language processing. *Neuroscience and Biobehavioral Reviews, 33*, 1168-1177.

APPENDIX A: POST-ERP QUESTIONNAIRE (CHINESE)

1、在刚才实验的过程中，您觉得有哪几种句子与图片含义不一致的情况？请举例并说明（请尽可能列举全面）。

2、请看下面的图片和句子。您认为这个句子是否和图片内容相一致？请解释你为什么这么认为。

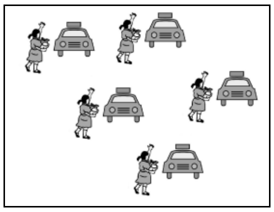


“图片里，有的女孩坐在毯子上。”

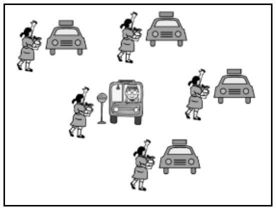
3、在下面的图片当中，请打勾来选择那些图片适合这个句子（可以多选）：

图片里，有的女孩在招出租车。

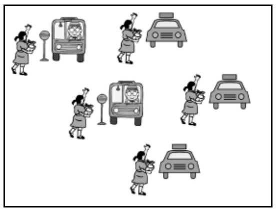
A）



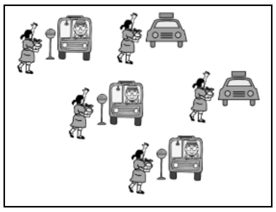
B）



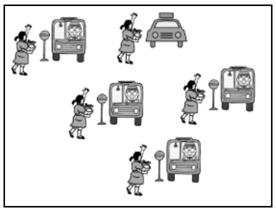
C）



D）



E）



4、在做实验的过程中，您有没有采用什么策略来决定是要看图的哪个部分？（例如：图片出现的时候，您预期下面听到的句子可能说的是图的哪个部分？）

5、请您认真阅读下面每一个句子，每个句子表达了一个命题，请判断这些命题的真假，1-假，7 代表真。1 到 7 分之间表示命题真假程度上的差别。

请您认真阅读下面每一个句子，并判断句子听上去是否别扭。请在下面的数字上画圈，1 代表十分别扭，7 代表十分正常。1-7 分之间表示程度上的差别。

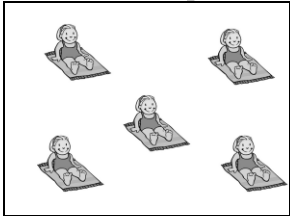
句子	真假（1为假、7为真）	别扭性（1为十分别扭、7为十分正常）
有的人有亲兄弟。	1 2 3 4 5 6 7	1 2 3 4 5 6 7
有的乌龟有贝壳。	1 2 3 4 5 6 7	1 2 3 4 5 6 7
有的上衣有扣子。	1 2 3 4 5 6 7	1 2 3 4 5 6 7
有的句子含有词语。	1 2 3 4 5 6 7	1 2 3 4 5 6 7
有的国旗上印有星星的图案。	1 2 3 4 5 6 7	1 2 3 4 5 6 7
有的大楼有电梯。	1 2 3 4 5 6 7	1 2 3 4 5 6 7
有的长颈鹿有脖子。	1 2 3 4 5 6 7	1 2 3 4 5 6 7
有的楼梯有台阶。	1 2 3 4 5 6 7	1 2 3 4 5 6 7
有的兔子长了耳朵。	1 2 3 4 5 6 7	1 2 3 4 5 6 7
有的公园有大树。	1 2 3 4 5 6 7	1 2 3 4 5 6 7

6、您认为这项实验的目的是什么？

APPENDIX B: POST-ERP QUESTIONNAIRE (ENGLISH TRANSLATION)

1. In the experiment, what types of picture-sentence inconsistencies did you notice? Please use examples and explain the inconsistency (please be as thorough as possible).

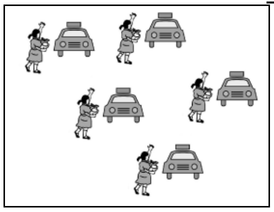
2. Please look at the picture and sentence below. Do you think this sentence is consistent with the content of the picture? Please explain why you think this way.

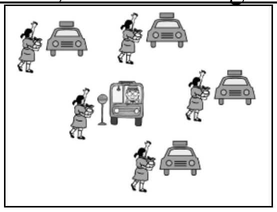


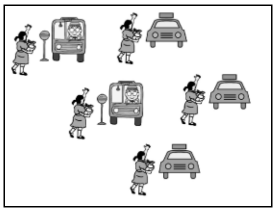
"In the picture, some of the girls are sitting on blankets."

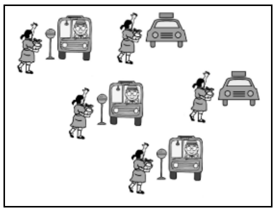
3. Among the following pictures, please indicate which pictures are consistent with the sentence. (You can choose more than one.)

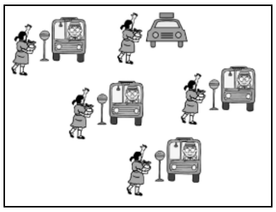
In the picture, some of the girls are hailing taxis.

A) 

B) 

C) 

D) 

E) 

4. While doing the experiment, did you use any particular strategy to decide which part of the picture to pay attention to? (For example: when the picture was shown, did you have any expectation about which part of the picture would be mentioned in the following sentence?)

5. Please carefully read each of the following sentences. Each sentence expresses a proposition. Please evaluate how true the expressions are (1 represents "false", 7 represents "true"). The numbers between 1 and 7 represent differences in the extent of truth or falsehood.

Please carefully read each of the following sentences and evaluate whether the sentence sounds awkward. Please chose one of the numbers below; 1 represents "very awkward", 7 represents "very normal", and the numbers in between represent differences in the extent of awkwardness or naturalness.

Sentence	Truth (1=false, 7=true)							Awkwardness (1=very awkward, 7=very natural)						
Some people have brothers.	1	2	3	4	5	6	7	1	2	3	4	5	6	7
Some turtles have shells.	1	2	3	4	5	6	7	1	2	3	4	5	6	7
Some shirts have buttons.	1	2	3	4	5	6	7	1	2	3	4	5	6	7
Some sentences have words.	1	2	3	4	5	6	7	1	2	3	4	5	6	7
Some flags have stars.	1	2	3	4	5	6	7	1	2	3	4	5	6	7
Some buildings have elevators.	1	2	3	4	5	6	7	1	2	3	4	5	6	7
Some giraffes have necks.	1	2	3	4	5	6	7	1	2	3	4	5	6	7
Some staircases have steps.	1	2	3	4	5	6	7	1	2	3	4	5	6	7
Some rabbits have ears.	1	2	3	4	5	6	7	1	2	3	4	5	6	7
Some gardens have trees.	1	2	3	4	5	6	7	1	2	3	4	5	6	7

6. What do you think the purpose of this experiment was?

APPENDIX C: SELF-PACED READING STIMULI (EXPERIMENTS 4 & 5)

Critical items

Item	Vignette	Comprehension question	Correct choice	Incorrect choice
1	Mary was preparing to throw a party for John's relatives. / She asked John whether (all of them/any of them) were staying in his apartment. / John said that (some of them/only some of them) were. / He added / that / the rest / would be / staying / in a hotel.	Who was Mary throwing a party for?	John's relatives	Her co-workers
2	Bill took out the fancy candles from the drawer. / He asked Claire whether (all of them/any of them) should be lit for dinner. / Claire said that (some of them/only some of them) should. / She added / that / the rest / would be / needed / later.	What were Bill and Claire lighting the candles for?	Dinner	A birthday cake
3	Susie heard Matthew's friends at the door. / She asked him whether (all of them/any of them) were going to the movie with him. / Matthew said that (some of them/only some of them) were. / He added / that / the rest / would be / too / busy.	Whose friends were at the door?	Matthew's	Susie's
4	The zookeeper was going to do a routine check-up on the lions. / He asked Robbie whether (all of them/any of them) had been fed that morning. / Robbie said that (some of them/only some of them) had. / He added / that / the rest / would be / fed / in the afternoon.	Who was going to check up on the lions?	The zookeeper	Robbie
5	In the store, Sally was unpacking the new shipment of shoes. / She asked Tiffany whether (all of them/any of them) should be marked on sale. / Tiffany said that (some of them/only some of them) should. / She added / that / the rest / would be / at / full price.	What was in the shipment?	Shoes	Pants
6	Mrs. Myers was worried that her students weren't ready for the test. / She asked the Mr. Robbins whether (all of them/any of them) had to take the test. /	Who was worried about their students?	Mrs. Myers	Mr. Robbins

	Mr. Robbins said that (some of them/only some of them) did. / He added / that / the rest / would be / able to / take it / the next weekend.			
7	Trevor was curious about the professors' summer breaks. / He asked Dr. Johnson whether (all of them/any of them) would be going out of town. / Dr. Johnson said that (some of them/only some of them) would. / He added / that / the rest / would be / around / all summer.	What break did Trevor ask about?	Summer break	Winter break
8	The local papers were all covering Jason's press conference. / Jason asked his publicist Sally whether (all of them/any of them) would run a front-page story. / Sally said that (some of them/only some of them) would. / She added / that / the rest / would be / running / other stories.	Where were some of the papers going to run the story?	The front page	The editorials section
9	Terry and his coach Rick were discussing top runners from the area. / Terry asked Rick whether (all of them/any of them) would be at the race. / Rick said that (some of them/only some of them) would. / He added / that / the rest / would be / resting / for the championships.	What's the coach's name?	Rick	Terry
10	Lydia was trying to choose one of the kittens from her friend Kim's pet store. / She asked Kim if (all of them/any of them) had gotten their shots yet. / Kim said that (some of them/only some of them) had. / She added / that / the rest / would be / getting them / later.	What was the mother concerned about?	Whether the kittens had shots	Whether the kittens were spayed
11	Justin wanted to find out which of Mike's desserts he could eat. / He asked Mike whether (all of them/any of them) could be made gluten free. / Mike said that (some of them/only some of them) could. / He added / that / the rest / would be / hard to do / that way.	What can't Justin eat?	Gluten	Dairy
12	Kurt and Brooke were thinking of selling their movies at a yard sale. / Brooke asked Kurt whether (all of them/any of them) were all right to sell. / Kurt said that (some of them/only	Why couldn't Kurt sell some of the movies?	Inappropriate for kids	Too unpopular

	some of them) were. / He added / that / the rest / would be / inappropriate / for kids.			
13	Jenna and Alex were making cupcakes. / Jenna asked Alex whether (all of them/any of them) were ready to be frosted. / Alex said that (some of them/only some of them) were. / He added / that / the rest / would be / ready / soon.	When would the cupcakes be ready?	Soon	Not for a while
14	Lyle was driving to meet his friends at the restaurant. / He asked Sarah whether (all of them/any of them) were already at the table. / Sarah said that (some of them/only some of them) were. / She added / that / the rest / would be / there / in five minutes.	When were the rest of the friends arriving?	Five minutes	Fifteen minutes
15	The students had prepared for the final presentation. / Allie asked Margaret whether (all of them/any of them) were supposed to present today. / Margaret said that (some of them/only some of them) were. / She added / that / the rest / would be / presenting / on Thursday.	When would the next day of presentations be?	Thursday	Monday
16	Darren knew his relatives would come for his birthday. / Darren asked his mother, Sally, whether (all of them/any of them) would be giving him clothes. / Sally said that (some of them/only some of them) would. / She added / that / the rest / would be / sending him / electronics.	What was Darren getting presents for?	Birthday	Christmas
17	Molly was looking at apartments. / She asked the agent Sally whether (all of them/any of them) would be available next month. / Sally said that (some of them/only some of them) would. / She added / that / the rest / would be / available / in August.			
18	Ryan had moved to a new town and was curious about the radio stations there. / He asked John whether (all of them/any of them) would play classic rock hits. / John said that (some of them/only some of them) would. / He added / that / the rest / would be / only / country music.			

19	The rock band was choosing cities to visit on their next tour. / They asked their manager Mary whether (all of them/any of them) would have good venues available. / Mary said that (some of them/only some of them) would. / She added / that / the rest / would be / booked / during the tour.			
20	Alex was perusing the breakfast cereals at the grocery store. / He asked Carrie whether (all of them/any of them) had high levels of sugar. / Carrie said that (some of them/only some of them) did. / She added / that / the rest / would be / healthier / but / not as tasty.			
21	The kids in the first grade class were getting antsy. / Kim asked Mrs. Brady whether (all of them/any of them) could go outside for recess. / Mrs. Brady said that (some of them/only some of them) could. / She added / that / the rest / would be / staying inside / to do / make-up work.			
22	The soccer players were all training very hard. / Eric asked Jack whether (all of them/any of them) would go to the tournament. / Jack said that (some of them/only some of them) would. / He added / that / the rest / would be / staying / in town.			
23	The clothing store just received a shipment of jeans. / Jared asked Erica whether (all of them/any of them) should be displayed out front. / Erica said that (some of them/only some of them) should. / She added / that / the rest / would be / kept / in the back / for now.			
24	Mrs. Landman was looking at laptops for her grandson. / She asked the employee Larry whether (all of them/any of them) were easy to carry around. / Larry said that (some of them/only some of them) were. / He added / that / the rest / would be / pretty heavy / to carry / to classes.			

25	Abby was taking her mom to meet her roommates. / Her mom asked Abby whether (all of them/any of them) would be home right now. / Abby said that (some of them/only some of them) would. / She added / that / the rest / would be / at work / or in class.			
26	Alexa was looking at the puppies at the shelter. / She asked the employee Stephen whether (all of them/any of them) would grow into large dogs. / Stephen said that (some of them/only some of them) would. / He added / that / the rest / would be / small / when full-grown.			
27	Eric was preparing to travel in Italy with his classmates. / Eric's mom asked him whether (all of them/any of them) had been to Europe before. / Eric said that (some of them/only some of them) had. / He added / that / the rest / would be / nervous about / traveling abroad.			
28	Carrie and Tim were comparing venues for their upcoming wedding. / Carrie asked Tim whether (all of them/any of them) would accommodate so many guests. / Tim said that (some of them/only some of them) would. / He added / that / the rest / would be / crowded / with / that many people.			
29	Lana was watching episodes of her favorite sitcom on her day off. / Ashley asked her whether (all of them/any of them) lasted less than an hour. / Lana said that (some of them/only some of them) did. / She added / that / the rest / would be / too rushed / that way.			
30	Arthur had set up tables for the garage sale. / Sarah asked him whether (all of them/any of them) would display the nice dishes. / Arthur said that (some of them/only some of them) would. / He added / that / the rest / would be / too / unstable.			
31	Brian had hired several new people to work at his restaurant. / The chef asked him whether (all of them/any of them)			

	would work in the kitchen. / Brian said that (some of them/only some of them) would. / He added / that / the rest / would be / servers / in the dining room.			
32	Josh wanted to talk about the new movie with his co-workers. / He asked his co-worker Seth whether (all of them/any of them) had seen the movie yet. / Seth said that (some of them/only some of them) had. / He added / that / the rest / would be / mad / if Josh / gave away spoilers.			
33	David and Brandon were looking at upcoming video games in the magazine. / David asked Brandon whether (all of them/any of them) would have a multiplayer mode. / Brandon said that (some of them/only some of them) would. / He added / that / the rest / would be / single player / only.			
34	Andrea was looking at computers in the store. / She asked the clerk Tom whether (all of them/any of them) came with webcams built in. / Tom said that (some of them/only some of them) did. / He added / that / the rest / would be / able to use / a USB webcam.			
35	Molly and Tony were hosting a dinner party for Tony's classmates. / Molly asked Tony whether (all of them/any of them) were allergic to any foods. / Tony said that (some of them/only some of them) were. / He added / that / the rest / would be / able to eat / anything.			
36	Max and Aaron were getting ready to host prospective students. / Max asked Aaron whether (all of them/any of them) needed rides from the airport. / Aaron said that (some of them/only some of them) did. / He added / that / the rest / would be / driving / themselves.			
37	Dr. Jones was scheduling end-of-semester meetings. / The secretary asked him whether (all of them/any of them) would be in the afternoon. / Dr. Jones said that (some of			

	them/only some of them) would. / He added / that / the rest / would be / early / in the morning.			
38	Tracy and Mark were trying to choose a restaurant downtown for their rehearsal dinner. / Mark asked Tracy whether (all of them/any of them) were open on the weekends. / Tracy said that (some of them/only some of them) were. / She added / that / the rest / would be / closed / at that time.			
39	Jason and Perry were ordering computers for the new lab. / Perry asked Jason whether (all of them/any of them) would have an Internet connection. / Jason said that (some of them/only some of them) would. / He added / that / the rest / would be / for offline work / only.			
40	Donna and Martha were discussing the rooms of the new house. / Donna asked Martha whether (all of them/any of them) would need to be wallpapered. / Martha said that (some of them/only some of them) would. / She added / that / the rest / would be / painted / instead.			
41	Brian had just finished checking the bikes in Tom's garage. / Tom asked Brian whether (all of them/any of them) needed to get a tune-up. / Brian said that (some of them/only some of them) did. / He added / that / the rest / would be / fine / for / another few months.			
42	Jim and Breanna were thinking of inviting their classmates to the movie theater. / Jim asked Breanna whether (all of them/any of them) were interested in artsy movies. / Breanna said that (some of them/only some of them) were. / She added / that / the rest / would be / bored / at those movies.			
43	Paul and Deb were trying to decide which gym to go to. / Paul asked Deb whether (all of them/any of them) had discounts for college students. / Deb said that (some of them/only some of them) did. / She added / that / the rest /			

	would be / at / full price.			
44	Greg was trying to choose which wine to order with dinner. / He asked the waitress Michelle whether (all of them/any of them) would go well with fish. / Michelle said that (some of them/only some of them) would. / She added / that / the rest / would be / too sweet / for that.			
45	Ed and Hillary were considering several universities for grad school. / Hillary asked Ed whether (all of them/any of them) were very competitive in admissions. / Ed said that (some of them/only some of them) were. / He added / that / the rest / would be / easy / to get into.			
46	Stan and Marilyn were trying to decide what type of tree to plant. / Stan asked Marilyn whether (all of them/any of them) would grow well in shade. / Marilyn said that (some of them/only some of them) would. / She added / that / the rest / would be / better off / in full sun.			
47	Quinn and Chase were thinking about going to one of the season's ball games. / Quinn asked Chase whether (all of them/any of them) were scheduled for the afternoon. / Chase said that (some of them/only some of them) were. / He added / that / the rest / would be / played / at night.			
48	Andy and Lisa were moving some pieces of furniture. / Lisa asked Andy whether (all of them/any of them) would fit in his car. / Andy said that (some of them/only some of them) would. / He added / that / the rest / would be / too big / for his car.			

Filler items

Fillers are grouped by typs/conditions. The naming convention for the types of filler trials is as follows. The first quantifier (before the hyphen) refers to the quantifier used in the context question (segment 2); the second quantifier (after the hyphen) refers to the quantifier used in the answer.

Vignette	**Comprehension question**	**Correct choice**	**Incorrect choice**
Filler type: all-some			
The student was concerned about the exams for this class. / He asked the teacher whether all of them had essay questions. / The teacher said that / some of them / did. / She added / that / he should / prepare carefully / beforehand.	What was the student concerned about?	The exams	The final project
Arnold was excited to meet the new students in the department. / He asked his professor whether all of them were out-of-state. / The professor said that / some of them / were. / She added / that / there were even / several / international students.	Were any new international students coming to the department?	Yes	No
Marty was trying to pick an ice cream flavor at the ice cream shop. / He asked the worker whether all of them were low-fat. / The worker said that / some of them / were. / He added / that / his favorite flavor / was / rainbow sherbet.	Who was getting ice cream?	Marty	Jim
Laurie was at the store on Black Friday looking at cameras. / She asked the clerk whether all of them were on sale. / The clerk said that / some of them / were. / He added / that / some / would also / come with / a free / memory card.	What did some of the cameras come with?	Memory card	Tote bag
Luke took the big bag of mushrooms out of the refrigerator. / He asked Jean whether all of them were going in the salad. / Jean said that / some of them / were. / She told Luke / to / pick out / the best ones / and give them / to her.			
Joshua and Kelly were trying to decide which caf?? to go to to do homework. / Joshua asked Kelly whether all of them had wireless internet. / Kelly said that / some of them / did. / She added / that / what was / more important / was whether / the coffee / was good.			
Heather was looking at new cars at the dealership. / She asked the salesman whether all of them had built-in GPS. / The salesman said that / some of them / did. / He added / that / it is / a really convenient / feature.			

Erik and Jonathan were grilling some steaks. / Erik asked Jonathan whether all of them should be well-done. / Jonathan said that / some of them / should. / He added / that / most people / are not / too picky.			
Bill and Diane had gathered logs for a fire. / Bill asked Diane whether all of them were dry enough to burn. / Diane said that / some of them / were. / She added / that / if they / needed / to gather more / they could.			
Noah and Eva were playing with toys in the yard. / Noah asked Eva whether all of them were waterproof. / Eva said that / some of them / were. / She added / that / they could / take them / in the pool.			
Doug and Sara were lighting sparklers. / Sara asked Doug whether all of them were multi-colored. / Doug said that / some of them / were. / He added / that / multi-colored / sparklers / are / the most / popular.			
Susie and Becky were looking at cottages to rent. / Susie asked Becky whether all of them were right on the lake. / Becky said that / some of them / were. / She added / that / those ones / were / the most / expensive.			
Filler type: any-some			
James was curious about Maggie's pet dogs. / He asked her whether any of them could do tricks. / Maggie said that / some of them / could. / She added / that / many of / their tricks / were / quite impressive.	Who had pet dogs?	Maggie	James
The coach was gathering up the volleyballs after practice. / He asked the players whether any of them were going flat. / The players said that / some of them / were. / They added / that / the nets / were also / a bit low.	How were the volleyball nets?	Too low	Just right
The teachers were chaperoning the kids on the school field trip. / Before leaving, Mr. Johnson asked whether any of them had asked to go to the bathroom. / Mrs. Baker said that / some of them / had. / She added / that / she / had / to go / too.	Where were the kids?	A field trip	Vacation
Teresa, a special education teacher, had a new group of students this year. / She asked her boss whether any of them had classroom assistants. / Her boss said that / some of them / did. / She added / that / the assistants / were all / highly trained.	What did Teresa teach?	Special education	Physical education
Grace and Joleen were trying to decide which of their friends to ask them to help move. / Joleen asked Grace whether any of them had a pickup truck. / Grace said that / some of them / did. / She added / that			

/ they should try / to think of / someone who / was not / very busy.			
Travis and Kim were trying to figure out which professor to take English from. / Travis asked Kim whether any of them gave study guides before exams. / Kim said that / some of them / did. / She added / that / she cared / more about / which professors / did not require / a final paper.			
Ian and Christine were trying to decide which yogurt shop to go to. / Christine asked Ian whether any of them had granola as a topping. / Ian said that / some of them / did. / He added / that / he wanted / granola / on his yogurt, / too.			
Adam and Johanna had been looking at new televisions. / Adam asked Johanna whether any of them would fit in their TV cabinet. / Johanna said that / some of them / would. / She added / that / they would / also need / to leave / space for / a DVD player.			
Lance and Olivia were thinking about visiting one of the museums in town. / Olivia asked Lance whether any of them had dinosaur exhibits. / Lance said that / some of them / did. / She added / that / they could / also / look for / a planetarium.			
Claire and Andrea were trying to decide which trivia team to join. / Claire asked Andrea whether any of them were any good. / Andrea said that / some of them / were. / She added / that / her old team / had won / the championship.			
The chef wanted to buy chicken from one of the local farms. / She asked her assistant whether any of them were certified organic. / The assistant said that / some of them / were. / He added / that / organic food / is becoming / more popular.			
Zach was looking at tomatoes at the farmer's market. / He asked his friend whether any of them looked ripe. / His friend said that / some of them / did. / He added / that / they should / buy / some basil, / too.			
Filler type: all-onlysome			
Ben wanted to learn about South American countries. / He asked Amy whether all of them were Spanish-speaking. / Amy said that / only some of them / were. / She added / that / many other / languages / are spoken / there, / too.	Which countries was Ben interested in?	South American countries	Asian countries
The little boy wanted to look at the books in the library. / He asked his mom whether all of them were picture books. / His mother said that / only some of	What did the boy want from the library?	Books	DVDs

them / were. / She added / that / the picture books / were / in the / kids' section.			
Dan was talking to his roommate Chen about people from Hong Kong. / Dan asked Chen whether all of them are bilingual in Chinese and English. / Chen said that / only some of them / are. / He added / that / it depends on / their age / and / education level.	Where is Dan's roommate from?	Hong Kong	Thailand
Sally was checking out books from the library. / She asked the librarian whether all of them were due the next month. / The librarian said that / only some of them / were. / She added / that / Sally / could check / the due dates / online.	How did the librarian say Sally could check the due dates?	Online	By phone
Jay was trying to choose which sandwich to get at the deli. / He asked whether all of them came with a soup. / The cashier said that / only some of them / did. / He added / that / the ones / that came / with soup / were indicated / on the menu.			
Sam was looking at the books on the list for his class. / He asked his friend whether all of them were required reading. / His friend said that / only some of them / were. / She added / that / Sam / could buy them / cheaper / online.			
Sonja and Alex were trying to decide which concert to go to. / Alex asked Sonja whether all of them were in the evening. / Sonja said that / only some of them / were. / She added / that / Alex / could choose / which one / to go to.			
Lee and Carol wanted to buy one of the hot tubs at the store. / Lee asked Carol whether all of them could sit eight people. / Carol said that / only some of them / could. / She added / that / she did not / foresee / needing / eight seats.			
Peter and Anne were considering presents for a baby shower. / Anne asked Peter whether all of them would work for either gender. / Peter said that / only some of them / would. / He added / that / she could / make / the final choice.			
Tyler and Rose wanted to serve ice cream to the guests. / Rose asked Tyler whether all of them could eat dairy. / Tyler said that / only some of them / could. / He added / that / sorbet / was also / available.			
Kylie and Lauren were trying to decide which air conditioner to buy. / Lauren asked Kylie whether all of them were environmentally friendly. / Kylie said that / only some of them / were. / She added / that / she cared / more / about / how powerful / they were.			

Carmen and Maria were picking flowers in the garden. / Maria asked Carmen whether all of them would last for several days in a vase. / Carmen said that / only some of them / would. / She added / that / they could / always / pick more.			
Filler type: any-onlysome			
The freshman was trying to decide which English class to take. / He asked his classmate whether any of them were particularly easy. / His classmate said that / only some of them / were. / He added / that / it depends on / the teacher.	What year was the student?	Freshman	Sophomore
The shopper was trying to decide which headphones to buy. / He asked the clerk whether any of them were sound-cancelling. / The clerk said that / only some of them / were. / He added / that / regular headphones / would be / just as good.	Did the clerk recommend sound-cancelling headphones?	No	Yes
Ellen was looking at the desserts at the buffet. / She asked the waiter whether any of them were sugar-free. / He said that / only some of them / were. / He added / that / the sugar-free / jello / was / really good.	Which dessert does the waiter recommend?	Sugar-free jello	Pudding
Will was trying to decide which tie to wear. / He asked Alice whether any of them went well with his suit. / She said that / only some of them / did. / She added / that / she / especially liked / the one / with / blue stripes.	Which tie did Alice like?	The blue striped one	The red one
Kristen and Ruth had to use one of the printers in the library. / Kristen asked Ruth whether any of them could print double-sided. / Ruth said that / only some of them / could. / She added / that / there might / be a long line / waiting / to use / those.			
Grant and Joel were looking at pumpkins in the pumpkin patch. / Joel asked Grant whether any of them were big enough to carve for Halloween. / Grant said that / only some of them / were. / He added / that / it was / a little early / to be thinking / about Halloween / anyway.			
Joey and Ryan wanted to eat one of the pies their mother was baking. / Ryan asked Joey whether any of them had chocolate in them. / Joey said that / only some of them / did. / He added / that / they / still needed / to bake / for a while.			
Sylvia and Audrey were trying to decide which color nail polish to use. / Sylvia asked Audrey whether any of them were sparkly. / Audrey said that / only some of them / were. / She added / that / the purple / was /			

an especially pretty / color.			
Donald and Shirley were looking at the cucumbers in the garden. / Donald asked Shirley whether any of them were big enough to pick. / Shirley said that / only some of them / were. / She added / that / the beans / were / in good shape, / though.			
Sophie and Jack had to choose which city to visit on vacation. / Jack asked Sophie whether any of them had a famous aquariu. / Sophie said that / only some of them / did. / She added / that / she / also wanted / to visit / a botanical garden.			
Jasper and Louise were flipping through channels on TV. / Louise asked Jasper whether any of them were showing a romantic comedy. / Jasper said that / only some of them / were. / He added / that / he / would prefer / to watch / a nature program.			
Neil and Eileen were trying to decide which team to support. / Eileen asked Neil whether any of them had cute mascots. / Neil said that / only some of them / did. / He added / that / he / did not / care much / about / mascots.			
Filler type: all-all			
Julie realized she had forgotten to put the chocolates in the fridge. / She called her roommate and asked whether all of them had melted already. / Her roommate said that / all of them / had. / She added / that / they / had made / a mess / on the counter.	Did the chocolates melt?	Yes	No
The secretary was collecting course evaluations. / She asked Lisa whether all of the evaluations had been completed. / Lisa said that / all of them / had. / She added / that / she had left / the pencils / there / for / the next section.	What did Lisa do with the pencils?	Brought them back	Left them in the room
Max and Jim needed to add carrots to their big salad. / Max asked Jim whether all of them had been chopped. / Jim said that / all of them / had. / He added / that / his hands / ached / from / all the chopping.	What was Jim chopping for the salad?	Carrots	Tomatoes
Sandi didn't want her cats running around during the dinner party. / She asked Ryan whether all of them had been shut in the upstairs room. / Ryan said that / all of them / had. / He added / that / he / had put / the litter box / there / as well.	Why did Sandi and Ryan put the cats upstairs?	Dinner party	Going on vacation
Bill and Jorge had to mark the roads the morning of the 5K race. / Bill asked Jorge whether all of them had been marked. / Jorge said that / all of them / had. / He added / that / he / had also / set up / the water stop.	What race were Bill and Jorge setting up?	5k race	Marathon
Jack was trying to choose which version of the	What kind of	Mac	PC

software to buy. / He asked the saleswoman whether all of them would run on a Mac. / The saleswoman said that / all of them / would. / She added / that / someone / could / help Jack / install / the software.	computer did Jack want to run the software on?		
Nikki was trying to convince Amy to read her favorite book series. / Amy asked Nikki whether all of them had happy endings. / Nikki said that / all of them / did. / She added / that / the stories / were / very interesting / as well.	What did Nikki like about the books?	The stories	The illustrations
Mitch was showing James the photos he had taken the other day. / James asked Mitch whether all of them had been touched up. / Mitch said that / all of them / had. / He added / that / it is / usually necessary / to touch up / the lighting / a little.	What does Mitch usually touch up?	The lighting	The colors
Terry was trying to choose a watch to buy. / He asked the clerk whether all of them had stopwatches. / The clerk said that / all of them / did. / He added / that / some / had / count-down timers / as well.			
Jackie and Rachel were talking about their co-workers. / Jackie asked Rachel whether all of them really rode their bikes to work. / Rachel said that / all of them / did. / She added / that / they were / always talking / about their bikes / around / the water cooler.			
Johnny's physical therapist had given him new exercises to do. / Johnny asked the physical therapist whether all of them were really necessary. / The physical therapist said that / all of them / were. / She added / that / they / would help / build up / his / back muscles.			
Jon was getting ready to take care of Kelly's cats for the weekend. / Jon asked Kelly whether all of them were outside cats. / Kelly said that / all of them / were. / She added / that / they still / always came / back in / when it was time / for food.			
The manager had many forms to sign. / He asked his secretary whether all of them were ready yet. / The secretary said that / all of them / were. / She added / that / they / had to / be finished / before / five o'clock.			
Paula wanted to try one of the new brands of cat food for her cat. / She asked the pet store employee whether all of them were high in protein. / The employee said that / all of them / were. / She added / that / the chicken-based / flavors / were / very popular.			
The teacher wanted to borrow his colleague's markers for the whiteboard. / He asked his colleague whether			

all of them worked. / His colleagues said that / all of them / did. / He added / that / there were also / more / in / the supply closet.			
Carla and Luke wanted to use one of the bright colors of paper to print flyers. / Luke asked Carla whether all of them were recyclable. / Carla said that / all of them / were. / She added / that / blue / was always / a good choice.			
Kristy and Jeff had picked strawberries at a farm. / Kristy asked Jeff whether all of them were gone. / Jeff said that / all of them / were. / He added / that / they / had been / delicious.			
Tim had watered the plants in the front yard. / His mother asked him whether all of them had actually needed watering. / Tim said that / all of them / had. / He added / that / the heat / was / really tough / on plants.			
Casey and Jody were washing dishes after dinner. / Jody asked Casey whether all of them could go in the dishwasher. / Casey said that / all of them / could. / He added / that / he / hated / washing dishes / by hand.			
Carrie was photocopying articles for her advisor. / She asked her advisor whether all of them were worth reading. / Her advisor said that / all of them / were. / She added / that / Carrie / would enjoy / them.			
Karen and Nick were getting the books from the reading list for their class. / Karen asked Nick whether all of them were in the university bookstore. / Nick said that / all of them / were. / He added / that / they / had both / new / and used / copies.			
Dustin was looking at comics at his friend's house. / He asked his friend whether all of them were classic editions. / His friend said that / all of them / were. / He added / that / they / were in / mint condition.			
Rick was sorting through the shoes in his closet. / He asked his son whether all of them were out of style. / His son said that / all of them / were. / He added / that / everything / his dad owned / was / out of / style.			
Richard and Cindy were baking potatoes for dinner. / Richard asked Cindy whether all of them were ready. / Cindy said that / all of them / were. / She added / that / the sour cream / was / already / on the table, / too.			
Filler type: any-all			
The kids were gathering up the game controllers to play video games. / They asked their mom whether any of them had batteries. / The mom said that / all of	What were the kids going to play?	Video games	Board games

them / did. / She added / that / they / could only / play for / thirty minutes.			
Anita was trying to choose one of the French textbooks. / She asked the clerk whether any of them came with CDs. / The clerk said that / all of them / did. / She added / that / the CDs / were / very helpful.	Who did Anita ask about the books?	The clerk	Her teacher
Lauren and Sally were trying to decide which mall in the city to go to on Saturday. / Lauren asked whether any of them had a good food court. / Sally said that / all of them / did. / She added / that / one / in particular / had / a great / pretzel stand.	Who was Sally going to the mall with?	Lauren	Bethany
Gary and Dana had picked up some snacks at the gas station. / Dana asked Gary whether any of them would spoil in the heat. / Gary said that / all of them / would. / He added / that / they would / be / home soon, / anyway.	Where did Dana and Gary get snacks?	The gas station	The mall
Derek and Sue were looking at rings at the jewelry store. / Derek asked Sue whether any of them caught her eye. / Sue said that / all of them / did. / She added / that / she / felt like / a kid / in a / candy store.	Did Julie like the rings?	Yes	No
Nathan wanted eat one of the sandwiches available in the cafeteria. / He asked his coworker whether any of them were decent. / His coworker said that / all of them / were. / She added / that / the daily special / was / a turkey BLT.	Where was Nathan planning on eating?	The cafeteria	A restaurant
Brad and Liz were looking at baby names in a book. / Liz asked Brad whether any of them were had historical significance. / He said that / all of them / did. / He added / that / some / were / more obscure / than / others.	What were Brad and Liz looking for names for?	A baby	A pet
Spencer was thinking about countries where he could go to teach English. / He asked his girlfriend whether any of them were appealing to her. / She said that / all of them / were. / She added / that / she / was really / up for / an adventure.	Why was Spencer going abroad?	To teach English	To study
Jordan was at the bookstore trying to choose a new poster to put on her wall. / She asked her roommate whether any of them would go with the colors in their room. / She said that / all of them / would. / She added / that / she / loved / the Monet print.			
Fred and Erin wanted to go running at one of the parks in town. / Fred asked Erin whether any of them had water fountains. / She said that / all of them / did. / She added / that / East Park / also had / restrooms.			
AJ and Ted wanted to try one of the new dishes at their favorite restaurant. / Ted asked AJ whether any			

of them sounded especially good. / AJ said that / all of them / did. / He added / that / he / did not know / which / to choose.			
Angela was very confused by the homework problems. / She asked Tim whether he had understood any of them. / Tim said that / he understood / all of them. / He added / that / he / would not mind / helping her / after class.			
Marisa and Colin wanted to buy a new showerhead at the hardware store. / Marisa asked Colin whether any of them had different pressure settings. / Colin said that / all of them / did. / He added / that / they / came in / metal / or plastic, / too.			
Kate was looking at the cat toys at the pet store. / She asked the salesman whether any of them had catnip in them. / He said that / all of them / did. / He added / that / the toy mice / were / very popular.			
Kathryn was trying to choose a picture book to read before bed. / She asked her big sister whether any of them were about animals. / Her big sister said that / all of them / were. / She added / that / her favorite / was / about / a dog.			
Lara and Joseph were considering which topping to get on their pizza. / Joseph asked Lara whether any of them would go well with green peppers. / Lara said that / all of them / would. / She added / that / she / felt like / mushrooms, / too.			
Heath and Steven were looking at new video game systems. / Steven asked Heath whether any of them were better than his old system. / Heath said that / all of them / were. / He added / that / the real issue / was / which one / had / the best games.			
Brendan and Everett had to pick which towels to bring to the beach. / Brendan asked Everett whether any of them were extra long. / Everett said that / all of them / were. / He added / that / they / should bring / a few / extras.			
Ernest and Nancy were trying to decide which buffet to go to. / Nancy asked Ernest whether any of them had a student discount. / Ernest said that / all of them / did. / He added / that / they / should choose / the one / that / was closest.			
Lola and Danny had picked up some avocados at the store. / Danny asked Lola whether any of them were ripe enough to eat. / Lola said that / all of them / were. / She added / that / they / should make / guacamole.			

Gabby was trying to decide which flavor of pudding to make for dessert. / She asked her father whether any of them sounded good to him. / He said that / all of them / did. / He added / that / he / would be / happiest / with / tapioca.			
Adrienne and Maddie wanted to take the canoe out on one of the lakes near their house. / Maddie asked Adrienne whether any of them had snapping turtles. / Adrienne said that / all of them / did. / She added / that / they would / be safe / in the canoe / regardless.			
Melissa wanted soft serve from one of the local ice cream shops. / She asked her mother whether any of them had sprinkles. / Her mother said that / all of them / did. / She added / that / she / liked / sprinkles, / too.			
Nico and has friends were trying to decide which of their cars to take to the concert. / Nico asked whether any of them had air conditioning. / His friend Pat said that / all of them / did. / He added / that / his own car / was / low / on gas.			
Filler type: noquantifier-otherquantifiers			
Stephanie loved the stones in Jim's rock collection. / She asked him whether they were from nearby. / Jim said that / many of them / were. / He added / that / the rest / he had / gotten / while traveling.	What did Jim have a collection of?	Stamps	Rocks
Anthony was thinking of joining his friend Tad's intramural soccer team. / Anthony asked Tad whether the players were very experienced. / Tad said that / many of them / were. / He added / that / the rest / were new / but / had already / improved a lot.	How were the new players in the team doing?	They had improved.	They had made no progress.
Jake and Charlie were talking about the bars in town. / Charlie asked whether they had happy hours. / Jake said that / many of them / did. / He added / that / the rest / had / various / other specials.			
Joe was on a college tour. / He asked the tour guide Gorden whether the classes there were discussion-based. / Gordon said that / many of them / were. / He added / that / the rest / were / labs or lectures, / but / very interesting.			
Hillary was visiting a college and was interested in the restaurants in town. / She asked an older student whether they used local suppliers. / The student said that / many of them / did. / He added / that / the rest / were / big chains.			
Jason and Jackie had a lot of library books to return. / Jackie asked whether they were renewable. / Jason			

said that / many of them / were. / He added / that / the rest / had / to be returned / right away, / though.			
Mark wanted to download songs from the band he had just heard. / He asked Alex whether their songs were on iTunes. / Alex said that / none of them / were. / He added / that / they / were all / on Youtube.	Where could Mark hear the band's songs?	Youtube	iTunes
Jackson was looking at sneakers in the shoe store. / He asked the clerk which brand made shoes with Velcro. / The clerk said that / none of them / did. / He added / that / Velcro / had gone / out of style.	Does the clerk think Velcro is popular nowadays?	No	Yes
Jan and Marcia were looking at blenders. / Marcia asked Jan whether they could make ice cream. / Jan said that / none of them / could. / She added / that / she / was trying / to avoid / dairy, / anyway.			
Nell was talking with her old soccer teammate, Emma, at the reunion. / She asked Emma whether her kids played soccer. / Ella said that / none of them / did. / She added / that / they / were on / the swim team.			
Cassie and Rich were looking at apples at the grocery store. / Cassie asked Rich whether the apples in this aisle were organic. / Rich said that / none of them / were. / He added / that / there was / an organic / aisle / around / the corner.			
John was looking at cell phones in the store. / He asked the clerk which ones had voice recognition. / The clerk said that / none of them / did. / He added / that / only / smart phones / have that.	What kinds of phones did the clerk say have voice recognition?	Smart phones	All phones
Michelle and her classmates wanted to take a group photo after dinner. / She asked her classmate Kenny who had a camera. / Kenny said that / none of them / did. / He added / that / they / would have / other chances / later.	Were the students able to take a picture?	No	Yes
Maureen and Bonnie wanted to go to one of the nearby beaches. / Maureen asked Bonnie whether they would have sharks. / Bonnie said that / none of them / would. / She added / that / shark attacks / are / profoundly rare, / anyway.			
Curtis and Joanne were looking at chandeliers. / Joanne asked Curtis whether they needed special lightbulbs. / Curtis said that / none of them / did. / He added / that / they / were / just like / normal lamps.			
Alice and Kara were planting oak trees. / Alice asked Kara whether they would need to be pruned regularly. / Kara said that / none of them / would. / She added /			

that / they / were very / low maintenance.			
Gabe and Marcus were hanging out after the first day of class. / Marcus asked Gabe whether the girls in his classes were single. / Gabe said that / none of them / were. / He added / that / he wouldn't / introduce / any girls / to Marcus, / anyway.			
Damian wanted a jacket with elbow patches. / He asked his father which stores in town sold jackets like that. / His father said that / none of them / did. / He added / that / he / could / order one / online.			
Beth and Emily were trying to decide which bar to go to. / Beth asked Emily whether the bars downtown were open until 2:00. / Emily said that / several of them / were. / She added / that / the rest / closed / at 1:00.	Are there bars open after midnight?	Yes	No
Ella was looking at the spices in the spice rack. / She asked Trey whether they were used in Indian food. / Trey said that / several of them / were. / He added / that / the rest / were / used in / other types / of food.	What was Ella looking at?	Spices	Drinks
Phil and Monica were looking at houses with the realtor. / Phil asked the realtor whether these houses had home security systems. / The realtor said that / several of them / did. / She added / that / the rest / could / have them / installed.			
Shannon and Joan were perusing the power tools in the garage. / Joan asked Shannon whether they were under warranty. / Shannon said that / several of them / were. / She added / that / the rest / were / too old.			
Hannah and Crystal wanted to buy wine glasses at the yard sale. / Crystal asked Hannah whether they were from a matched set. / Hannah said that / several of them / were. / She added / that / the rest / were / unique.			
Geoff and Marion were trying to decide which island to visit. / Geoff asked Marion whether they were close enough for a day trip. / Marion said that / several of them / were. / She added / that / the rest / were / a bit / farther away.			
Teresa had shared a cookie recipe with Greg. / She asked Greg whether the cookies had turned out all right. / Greg said that / most of them / had. / He added / that / the rest / had gotten / a little / burnt.	What happened to some of the cookies?	They got burnt.	They broke.
Scott was getting ready to order new software in the library. / He asked the technician whether the computers were compatible. / The technician said that / most of them / were. / He added / that / the rest /	What was Scott ordering for the library?	Software	Books

would / have to be / upgraded / first.			
Elaine and Cliff were at a wedding reception. / Elaine asked Cliff whether the drinks there were non-alcoholic. / Cliff said that / most of them / were. / He added / that / the rest / were / alcoholic.			
Jerry and Sheila wanted to bring gum on the airplane. / Sheila asked Jerry whether the ones at the newsstand were sugar-free. / Jerry said that / most of them / were. / He added / that / the rest / did not / look / very good.			
Gerry and his son were at the amusement park. / Gerry asked the attendant whether the rides had height limits. / The attendant said that / most of them / did. / He added / that / the rest / were still / fun.			
Alan and Frank wanted to visit the museum. / Alan asked Frank whether the new exhibits were open to the public. / Frank said that / most of them / were. / He added / that / the rest / would be / open / soon.			
Amy was impressed by the other squash players in the club she was joining. / She asked another player whether they had had coaching. / The player said that / a few of them / had. / She added / that / the rest / had just / picked it up / over time.	What club was Amy joining?	Squash club	Chess club
Joshua felt that the computers in the library were too slow. / He asked Seth whether the computers ran slowly for him too. / Seth said that / a few of them / did. / He added / that / the rest / he / had not tried / yet.	Where were Joshua and Seth testing computers?	The library	The student union
Dylan and Janelle were going to look at apartments. / Janelle asked Dylan whether the units downtown had washing machines. / Dylan said that / a few of them / did. / He added / that / the rest / were / near / the laundromat.			
Jeremy and Brett were planning a hike. / Jeremy asked Brett whether the trails on this side of town were hilly. / Brett said that / a few of them / were. / He added / that / the rest / were / poorly maintained, / though.			
Colleen and Edward wanted to go bowling. / Edward asked Colleen whether the bowling alleys in town had weeknight specials. / Colleen said that / a few of them / did. / She added / that / the rest / had / better food.			
Liza and Beverly had just picked their classes. / Liza asked Beverly whether her classes had labs. / Beverly said that / a few of them / did. / She added / that / the rest / were / language / classes.			
The professors were discussing some students'	Who had given	Students	Job

presentations. / Dr. Smith asked Dr. Rivera whether the presentations had too many animations. / Dr. Rivera said that / two of them / did. / He added / that / the rest / were / all right.	presentations?		applicants
Glenn and Leah were planning to go to one of the concerts in town. / Glenn asked Leah whether they would have a light show. / Leah said that / two of them / did. / She added / that / the rest / were / more / low-key.	Who was Glenn going to a concert with?	Leah	Marty
Natalie and Dean wanted to take a train on their vacation. / Natalie asked Dean whether they would have dining cars. / Dean said that / two of them / did. / He added / that / the rest / were / commuter trains.			
Tess and Wayne were shopping for a new camera. / Tess asked Wayne whether the models in this store had telephoto lenses. / Wayne said that / three of them / did. / He added / that / the rest / were / point-and-clicks.			
Chelsea and Wyatt were talking about going to a movie. / Wyatt asked Chelsea whether the movies in the theater were in 3-D. / Chelsea said that / three of them / were. / She added / that / the rest / were / normal.	Where did they want to see a movie?	In the theater	At home
Josie and Christian were looking at the chickens in the coop. / Josie asked Christian whether the chickens were old enough to lay eggs. / Christian said that / three of them / were. / He added / that / the rest / would be / ready / soon.			
Jessie and Caleb were getting ready to go to orientation. / Caleb asked Jessie whether the workshops at orientation were mandatory. / Jessie said that / two of them / were. / She added / that / the rest / were / optional / but / recommended.			
Chad and Tori wanted to adopt a dog. / Chad asked Tori whether the dogs at the pound were fixed. / Tori said that / three of them / were. / She added / that / the rest / weren't / old enough / yet.			
Drew and Giles were in the computer lab. / Giles asked Drew whether the documents had printed correctly. / Drew said that / four of them / had. / He added / that / the rest / had gotten / jammed.	What had happened to the documents?	Got jammed	Out of paper
Mandy and Shelby wanted to go out to eat. / Mandy asked Shelby whether the restaurants downtown served brunch. / Shelby said that / four of them / did. / She added / that / the rest / had / great / breakfast / options.			

Damon was doing laundry. / He asked his mother whether the shirts could go in the dryer. / His mother said that / four of them / could. / She added / that / the rest / should be / hung up / to dry.			
Ethan and Connor had just seen a big crash in the bike race. / Ethan asked Connor whether those riders were still in the race. / Connor said that / four of them / were. / He added / that / the rest / had / dropped out.			

Practice items

Vignette	**Comprehension question**	**Correct choice**	**Incorrect choice**
Janice wanted to go for a bike ride on the weekend. / Her friend Cathering was going to go with her. / They decided / to go / on Saturday. / On Saturday / they / packed up / a picnic lunch / and set out.	When did Janice and Catherine go biking?	Saturday	Sunday
Johnny and Rich were supposed to bring dessert to the party. / They decided to bring ice cream, but couldn't choose a flavor. / Rich's favorite flavor / was / cookies and cream. / Johnny's / favorite flavor, / however, / was / rocky road.	What were Johnny and Rich bringing to the party?	Dessert	Appetizers
Peter and Sam were both dog lovers. / Peter asked Sam what his favorite breed was. / Sam said / his favorite was / beagles. / He added / that / they made / great / house pets.			
Clint and Ted were in the same chemistry class. / Clint asked Ted when their final was. / Ted said that / it was / next Wednesday. / He added / that / there was / a final paper / due / on Friday, / also.			
Jeff and Mark had to turn in their class project on Monday. / Jeff asked Marke to send him his part by Saturday night. / Mark nodded / and said that / he would. / He added / that / he would / proofread it / first.			
The dinosaurs went extinct 65 million years ago. / Scientists are still not sure what caused the extinction. / Some believe / it was caused by / a meteor. / Others believe / it was / due to / volcanic activity / or a / sudden drop in / sea level.			

APPENDIX D: STROOP TASK STIMULI

Word reading	Color naming	Incongruent
YELLOW	XXXX	RED
BLUE	XXXX	RED
BLUE	XXXX	YELLOW
RED	XXXX	BLUE
RED	XXXX	GREEN
BLUE	XXXX	GREEN
BLUE	XXXX	YELLOW
BLUE	XXXX	RED
BLUE	XXXX	YELLOW
RED	XXXX	BLUE
BLUE	XXXX	RED
RED	XXXX	BLUE
YELLOW	XXXX	BLUE
YELLOW	XXXX	GREEN
GREEN	XXXX	BLUE
BLUE	XXXX	GREEN
GREEN	XXXX	BLUE
BLUE	XXXX	GREEN
RED	XXXX	BLUE
BLUE	XXXX	YELLOW
BLUE	XXXX	RED
RED	XXXX	YELLOW
YELLOW	XXXX	BLUE
GREEN	XXXX	BLUE
YELLOW	XXXX	BLUE
YELLOW	XXXX	GREEN
BLUE	XXXX	YELLOW
BLUE	XXXX	RED
GREEN	XXXX	RED
RED	XXXX	GREEN
YELLOW	XXXX	RED
BLUE	XXXX	RED
BLUE	XXXX	RED
BLUE	XXXX	RED
YELLOW	XXXX	BLUE
YELLOW	XXXX	BLUE
GREEN	XXXX	YELLOW
GREEN	XXXX	YELLOW
RED	XXXX	GREEN
RED	XXXX	BLUE

RED	XXXX	YELLOW
BLUE	XXXX	RED
BLUE	XXXX	YELLOW
YELLOW	XXXX	GREEN
BLUE	XXXX	RED
YELLOW	XXXX	GREEN
YELLOW	XXXX	BLUE
GREEN	XXXX	YELLOW
YELLOW	XXXX	GREEN
YELLOW	XXXX	GREEN
GREEN	XXXX	YELLOW
RED	XXXX	GREEN
YELLOW	XXXX	GREEN
GREEN	XXXX	RED
RED	XXXX	GREEN
YELLOW	XXXX	BLUE
BLUE	XXXX	GREEN
RED	XXXX	YELLOW
GREEN	XXXX	RED
YELLOW	XXXX	RED
RED	XXXX	YELLOW
YELLOW	XXXX	BLUE
GREEN	XXXX	YELLOW
GREEN	XXXX	BLUE
YELLOW	XXXX	GREEN
BLUE	XXXX	YELLOW
GREEN	XXXX	RED
GREEN	XXXX	BLUE
YELLOW	XXXX	BLUE
GREEN	XXXX	RED
RED	XXXX	BLUE
GREEN	XXXX	RED
BLUE	XXXX	GREEN
YELLOW	XXXX	BLUE
GREEN	XXXX	RED
RED	XXXX	GREEN
BLUE	XXXX	GREEN
GREEN	XXXX	YELLOW
RED	XXXX	GREEN
YELLOW	XXXX	GREEN
RED	XXXX	GREEN
GREEN	XXXX	BLUE
GREEN	XXXX	YELLOW

BLUE	XXXX	GREEN
RED	XXXX	YELLOW
GREEN	XXXX	BLUE
RED	XXXX	YELLOW
GREEN	XXXX	RED
RED	XXXX	GREEN
RED	XXXX	YELLOW
BLUE	XXXX	RED
YELLOW	XXXX	RED
RED	XXXX	GREEN
GREEN	XXXX	BLUE
RED	XXXX	YELLOW
YELLOW	XXXX	RED
RED	XXXX	YELLOW
YELLOW	XXXX	BLUE
GREEN	XXXX	RED
GREEN	XXXX	YELLOW

APPENDIX E: INDIVIDUAL DIFFERENCES DESCRIPTIVE STATISTICS

Descriptive statistics for the individual difference measures, both aggregated across experiments and broken down by experiment. These descriptives were conducted prior to any transformations and normalizations described in the text. Further information on each of these measures is as follows:

- **Stroop**: This is the size of the Stroop effect (completion time for the Incongruent color naming set minus completion time for the Congruent color naming set). The unit of measurement is seconds.
- **AQ measures**: These are the scores on the five Autism-Spectrum Quotient subscales. Each scale has a minimum score of 0 and a maximum score of 10.
- **IRI measures**: These are scores on the four Interpersonal Reactivity Index subscales. Each scale has a minimum score of 0 and a maximum score of 28.
- **Truth rating**: This is the average truth rating participants gave to underinformative sentences; the minimum rating on this scale is 1 (very untrue) and the maximum is 7 (very true).
- **Reading Span and Counting Span measures**: Recall is the percentage of trials recalled correctly. This measure was not used in the data analyses reported in the text, but is reported here because it is on a meaningful scale. The analyses reported in the text were based on composite scores (see the text for explanation of how composite scores were computed), and because the raw scores were sphered in order to calculate composite scores, the composite scores by necessity roughly approximate a standard normal distribution.

- **Flanker**: This is the size of the flanker effect (response time for Incongruent trials minus response time for Congruent trials). The unit of measurement is milliseconds, and the effect is not a raw mean difference but is a regression coefficient (see the text for more details about the regression model used to calculate flanker effects). Note that for the analyses presented in the text, reaction times were transformed via reflected reciprocal prior to computing flanker effects; here the flanker effects are reported based on raw reaction times instead, in order to show the effects on a more meaningful scale.

Experiments 4 & 5

	Mean	Median	SD	Range	Skewness
Stroop	17.79	16.65	7.99	-0.37–40.39	0.72
AQ Social Skill	2.44	2	2.39	0–9	1.05
AQ Attention Switching	4.82	5	1.81	1–9	0.13
AQ Attention to Detail	5.89	6	2.23	0–10	-0.27
AQ Communication	2.12	2	1.71	0–6	0.58
AQ Imagination	2.45	2	1.6	0–8	0.91
IRI Perspective	15.97	16	4.2	4–24	-0.55
IRI Fantasy	19.88	21	5.25	6–28	-0.71
IRI Empathy	20.97	21.5	4.65	6–27	-0.84
IRI Distress	11.65	11.5	4.67	2–22	0.48
Truth Rating	4.97	6.1	2.25	1–7	-0.61
Reading Span (Recall)	64.62	67.5	19.9	0–97.91	-1.25
Reading Span (Composite)	0	0.13	0.78	-3.08–1.29	-0.98
Counting Span (Recall)	61.24	62.17	16.76	8.89–96.44	-0.58
Counting Span (Composite)	0	0.14	0.8	-2.37–1.45	-1.09
Flanker	108.56	109.16	46.76	-32.77–269.21	0.22

Experiment 4

	Mean	Median	SD	Range	Skewness
Stroop	17.4	16.59	6.78	8.75–34.17	1.1
AQ Social Skill	1.96	1.5	1.67	0–6	0.56
AQ Attention Switching	4.71	4	2.02	1–9	0.33
AQ Attention to Detail	5.82	6	2.09	2–10	-0.17
AQ Communication	1.82	1	1.56	0–6	1.24
AQ Imagination	2.32	2	1.44	0–5	0.52
IRI Perspective	15.25	14.5	3.46	8–22	0.04
IRI Fantasy	19.89	20.5	4.78	11–28	-0.25
IRI Empathy	20.32	20.5	4.6	9–27	-0.32
IRI Distress	11.25	11	3.99	4–19	-0.06
Truth Rating	4.73	4.8	2.45	1–7	-0.43
Reading Span (Recall)	70.02	70.69	15.16	41.53–97.92	-0.07
Reading Span (Composite)	0.09	0.29	0.67	-1.11–1.19	-0.1
Counting Span (Recall)	64.82	65.72	16.84	20–96.44	-0.6
Counting Span (Composite)	0	0.1	0.82	-2.1–1.21	-0.6
Flanker	106.08	105.01	52.35	25.65–269.21	0.89

Experiment 5

	Mean	Median	SD	Range	Skewness
	Mean	Median	SD	Range	Skewness
Stroop	18.08	16.84	8.85	-0.37–40.39	0.52
AQ Social Skill	2.86	2	2.78	0–9	0.81
AQ Attention Switching	4.89	5	1.7	1–8	-0.07
AQ Attention to Detail	5.97	6	2.39	0–10	-0.35
AQ Communication	2.41	2	1.77	0–6	0.15
AQ Imagination	2.59	2	1.72	0–8	0.97
IRI Perspective	16.81	17	4.28	4–24	-0.81
IRI Fantasy	19.84	21	5.71	6–27	-0.85
IRI Empathy	21.43	23	4.75	6–27	-1.18
IRI Distress	11.68	13	4.93	2–22	-0.05
Truth Rating	5.09	6.2	2.12	1–7	-0.67
Reading Span (Recall)	61.27	65.69	22.18	0–88.89	-1.35
Reading Span (Composite)	-0.07	0.08	0.87	-3.08–1.29	-1.13
Counting Span (Recall)	58.69	60.28	16.65	8.89–90.22	-0.63
Counting Span (Composite)	0.02	0.15	0.81	-2.37–1.45	-1.47

APPENDIX F: BACKGROUND SPEECH STIMULI

Real words

Monday
traffic
stronger
economic
value
go
delegate
rugby
years
consumer
scheduled
off
eminent
detours
open
overtime
him
confirm
gained
against
impressive
bill
appointment
extremely
agreed
only
breakfast
appeal
behind
cookies
house
ever
class
visited
news
hall
eighteenth
cutting
commissions
officer's
exhibition
companies
special
science
half
boyfriend
driving
residents
delivered
today
weather
fold
wind
dinner
center
renovate
artists
ago
unit
city
traditions
are
historians
later
given
quarterfinal
races
urged
work
know
construction
cross-town
after
shredded
recycled
daisy
children
presentation
never
scored
young
unanimous
sketches
sun
site
other
unexpectedly
attempt
began
developer
nobody
contains
exceptional
warmly
payment
windy
ruined
recovered
writing
dark
official
highway
night
race
goals
out
themed
light
neighbour
pretty
she
starters
take
action
freeze
finally
when
everyone
involved
sunrise
told
aluminum
bond
expected
flowers
nursery
maternity
goal
students
wrapped
Kansas
we're
capsule

team
usually
daily
photocopy
took
local
paintings
options
fast
hope
winter
percent
streets
hours
rainfall
soccer
mothers
street
community
close
move
early
into
approved
lodge
including
educational
friends
has
back
state
for
war
will
warm
regional
three
neighborhoods
championship
process
products
elementary
newspaper
going
searching
aims
tissue

concert
served
plan
station
European
day
thought
Chinese
persistently
present
finding
old
twilight
committee
costumed
westbound
attracted
trick-or-treat
candidates
positive
increasing
six
earthquake
nowadays
being
seems
more
entirely
warmth
used
said
cheese
month
many
advised
all
shined
raid
intersection
unions
several
fifths
looking
almost
spaces
hurricane
shirtless

could
newsprint
of
got
lot
mentioned
green
five
speech
traveler
entire
civil
asked
teammate
famous
their
meet
pellets
vacant
life
away
outcry
completely
coffee
started
homecoming
what
edge
opera
redevelopment
place
collect
visit
following
art
cover
roundabout
valleys
time
concerns
monetary
side
ground
Mister
tremendous
went
new

macaroni
lanes
recommend
noticed
university
shut
Sally
urban
called
caused
influence
currently
get
shoofly
encourage
which
faulty
cloak
throughout
bloomed
expressing
unable
residential
March
convenient
nearly
seven
immediately
capital
court
requires
now
programming
small
first
recommendation
younger
will
amount
should
coaches
political
homecoming
disputing
recently
already
difficult

bridge
program
November
again
product
surgery
always
universe
climate
car
set
opponent
cold
need
progress
gave
fiberfill
board
reform
quietly
remains
great
accurate
banks
even
minutes
begin
travel
forever
money
reason
October
schools
this
and
additional
hear
came
among
shop
confess
will
win
found
lane
printing
homecoming

location
will
gallery
man
problems
considered
severe
porch
make
latest
giving
good
previous
opened
from
succeed
society
average
without
line
seek
plastic
integrity
force
turn
welcome
instead
swim
courthouse
seeking
next
party
eastbound
market
campus
take
corroboration
predicting
buildings
involve
saw
can
soldiers
generally
approached
strange
area

game
project
bed
football
quarter
didn't
important
with
room
paper
traffic
making
classrooms
medieval
having
characters
brother
were
ways
like
eight
agreements
steps
along
featured
picture
week
busy
crossed
much
find
long
agreement
equations
important
focus
there
hard
places
succeeded
installed
technicians
train
urging
his
varsity
acquired

Tuesday
its
over
parents
sold
mystery
least
members
woman
trade
play
takes
solution
Missouri
earlier
promises
will
field
through
Friday
middle
advance
claim
person
locations
queen
voter
needs
blew
planning
had
best
library
government
pray
unmarked
press
big
thought
mid-1900s
exciting
afternoon
right
parking
finished
fight
four

television
boy
they
spite
high
made
fine
international
events
hood
still
will
recent
rivals
approving
order
than
clear
paperboard
cider
nine
come
century
slipped
have
one
significant
any
debate
square
produce
success
required
wasn't
past
commissioners
o'clock
support
dated
will
robin
become
moved
rivalry
tradition
little
higher

that
well
loans
sleeping
stolen
charming
bag
part
easiest
hike
invited
council
just
museum
closely
player
energy
home
each
didn't
public
parallel
central
includes
school
parade
announced
coach
important
supermarket
recycling
late-October
during
technical
grand-parent's
built
yet
north
jackets
producing
introduced
most
around
transported
few
improvements
about

better
lost
leave
season
job
homecoming
been
copy
although
actual
getting
changes
defensively
will
increased
until
left
last
very
obliged
doughnuts
western
meetings
people
crowning
birth
played
various
strong
quite
win
hits
rebuilding
every
group
watch
ten
voting
tied
supporters
scheming
tenth
nephew
mixed
student
damage
onto

decided
production
specialist
easy
homecoming
our
year
older
two-year-old
mischief
then
twice
family's
burial
items
town
will
club
killed
main
case
determined
ski
problem
settlement
election
village
soon
needed
teacher
materials
scabbard
ballast
mastiff
tomahawk
suburban
emanation
platypus
felicity
competent
neptune
platform
emigration
tapestry
velveteen
sturgeon
bungalow

glisten
grimace
abdomen
nectar
herbivore
frontier
sovereign
padlock
sailboat
classroom
rattlesnake
drainpipe
cornfield
stopwatch
mousetrap
bathrobe
handgun
newspaper
videotape
sandstorm
airplane
nosedive
seatbelt
hometown
teacup
backbone
daydream
headache
toothpaste
bubblegum
paintbrush
cellphone
beefsteak
flagpole
treetop
doorknob
basketball
hairspray
chainsaw
spacecraft
forklift
whiplash
grapefruit
earthquake
frostbite
truckload
shoebox
codename
reindeer
flashbulb
postcard
worldview
folklore
walkman
fingertip
beanstalk
milkshake
humankind
breakfast
ashtray
cutthroat
thunderbolt
heartburn
tombstone
earplug
standpoint
southwest
landlord
payroll
tailgate
hogwash
dashboard
crackpot
passport
bottleneck
honeymoon
eggplant
jailbird
doughnut
sugarcane
storefront
rainbow
brainchild
rollerblade
hamstring
windfall
turncoat
bootleg
bookworm
armpit
hallmark
warpath
bombshell
pineapple
bandwagon
doghouse
bedrock
peppermint
sherbet

Novel words

sunkay
trammick
strenger
tecopomic
zaluke
vo
gelegote
sugby
kears
tonsamer
beduted
oss
ummifent
dapours
ipent
opersime
lim
gonsirm
gined
pʌgets
limtressive
fiss
sottoimpment
dexgreesely
hegreef
pownzie
kressfast
aggeaf
tee-pined
noosies
dauf
iveck
sliss
kuzzitted
zoof
kell
sate beenth
gittick
tiggissions

luffickers
jexquitition
gambanies
skemmle
tiants
kaff
loytremp
scriping
cressipents
seloovered
moofay
plecker
tolk
slint
lissle
manter
senosate
kustists
seefo
byune-mit
liffy
tromitions
frass
huspories
yaker
slissen
warfer linel
yajes
wurged
ferk
klow
spink
sunkriction
kosspow
afkit
sneffid
befikled
faipy
milren
reskimpation
gepper
skappid
nup
ucrisafous
sletchid
spone
geep

ahthick
mundebekt
magint
bofan
kromeleper
bomuddy
suftark
kovarkinal
grengly
plactent
zovty
lendack
provovvered
vugern
krawble
zafissel
zeemay
gige
jerrip
jollid
prouk
thoked
zipe
glaybour
riggip
hosh
stoofels
klipe
spaction
griesel
fopally
gemmed
keffry-kunn
binkolled
lupripe
glolk
adoofameer
klonnit
pextected
kluffers
burspery
tagrenity
jopal
tubrents
gulloped
sankoon
kreer

sapkool
trome
blupally
klagey
voco-tobby
koop
flople
saimpings
smopkins
plusk
zope
fintle
terpenk
treefs
scowper
fainrall
kosser
thommers
treef
makkonity
slose
voose
larley
tagone
sarooved
chodge
clinsooding
capsamational
dreffs
sazz
bock
fleek
froo
clorp
leewo
rama
geerinal
pleef
begoroods
panctionpip
rospess
krompus
ploctement
skoozvaver
goptink
ferchunt
smase

smickle
sonkert
verked
palmin
musion
purowegan
vok
potht
supese
poskispently
seevent
kinding
drog
kwipight
mikippity
soctumed
bowstwend
taracted
kroperdeet
dankifets
soppessive
ombreasing
hidgel
gerdnack
nupalaise
leemick
freems
klore
tensipely
marwith
jumie
hinghud
treamnug
prieldlisk
scringhud
stropror
mellcond
fampror
tarknane
vaithaib
plarmbip
sooftarg
eronkie
mourdgid
dacklarch
brumphom
flifflounе

fextcralk
fubrol
nangak
glormpeb
fudral
segrask
speeldorp
plewnofe
pliptond
siblusp
gorbux
dirser
stacha
niehan
higpoy
veyjun
daketrel
lirgtorg
chawmord
brolchon
foshtule
slepbort
pisknert
sumemirt
himepron
spowtenk
tritferd
chudhake
kerdhaif
shisebisp
lidefalb
malshplich
frokeskeer
thichprip
cradefodge
slentgoost
plourtrenk
spreljeash
plachsork
cralytroud
prundgrall
grenfelgraple
gorbnom
dirnoy
stadaw
nieyeg
higtig

veylod
dakeblar
lirgfoll
chawpraw
brolsare
foshstit
slepnorn
piskforb
sumesnel
himestib
spowfler
tritfeep
chudleem
kerdplip
shiseferk
lidesleg
malshhaist
frokelaip
thichtagar
cradethean
slentfrand
plourmarpy
sprelplard
plachtronk
cralypreed
prundchish
grenfelslempor
nomux
noyser
dawcha
yeghan
tigpoy
lodjun
blartrel
folltorg
prawmord
sarechon
stittule
nornbort
forbnert
snelmirt
stibpron
flertenk
feepferd
leemhake
pliphaif
ferkbisp

slegfalb
haistplich
laipskeer
tagarprilb
theanfodge
frandgoost
marpytrenk
plardjeash
tronksork
preedtroud
chishgrall
slemporgraple
nomgorb
noydir
dawsta
yegnie
tighig
lodvey
blardake
folllirg
prawchaw
sarebrol
stitfosh
nornslep
forbpisk
snelsume
stibhime
flerspow
feeptrit
leemchud
plipkerd
ferkshise
sleglide
haistmalsh
laipfroke
tagarthich
theancrade
frandslent
marpyplour
plardsprel
tronkplach
preedcraly
chishprund
slemporgrenfel
uxgorb
serdir
chasta

hannie
poyhig
junvey
treldake
torglirg
mordchaw
chonbrol
tulefosh
bortslep
nertpisk
mirtsume
pronhime
tenkspow
ferdtrit
hakechud
haifkerd
bispshise
falblide
plichmalsh
skeerfroke
prilbthich
fodgecrade
gooststlent
trenkplour
jeashsprel
sorkplach
troudcraly
grallprund
graplegrenfel
uxnom
sernoy
chadaw
hanyeg
poytig
junlod
trelblar
torgfoll
mordpraw
chonsare
tulestit
bortnorn
nertforb
mirtsnel
pronstib
tenkfler
ferdfeep
hakeleem

haifplip
bispferk
falbsleg
plichhaist
skeerlaip
prilbtagar
fodgethean
goostfrand
trenkmarpy
jeashplard
sorktronk
troudpreed
grallchish
grapleslempor
glid
pyleg
otlat
ock
hed
spale
lant
elt
reaken
midbith
rint
uit
ong
rog
pring
ress
ird
oon
ane
ip
gerflibble
eld
oard
tage
troke
tyle
lemind
ront
rewnie
prock
haft
ther
tring

ealt
crod
rass
ber
hing
ast
pebug
elon
hape
stide
paxe
yapple
orse
rame
rop
creen
chool
netap
spea
hild
nake
acken
oach
erm
snut
trett
tecar
sook
kenk
senk
fint
godie
nank
yube
surt
varn
rost
fote
isel
mimp
houb
slig
aren
hade
argle
drope
lorse

yean
bafe
onch
plort
trug
prew
nable
torb
solt
lork
hane
eart
feag
trone
flet
blenk
slorm
marel
phenk
gath
plest
nume
shenk
smey
mople
steg
taple
gine
morsh
negle
moit
lote
morch
sife
brup
lorp
proke
mertesh
sorge
fenth
nork
meast
vome
dast
baprel
vogtist
yitfane

nilkad
blapdum
glospum
forpmerk
pridnusk
treepshorm
merbtarn
flimhan
fopshreen
jerglem
filbreng
greldem
gredmanch
fipslen
larfbast
hasemisk
prinkow
falphort
hargpilt
deneskine
frageclest
brimesheme
flainchenk
slapetrosh
jaiseclim
bramabome
brindnorg
brendfreem
flebarganch
fomclem
fiskpap
detwose
lupfrant
pabhest
daltimp
blashlask
fleptrud
pristrem
feshmorp
feskprap
fingtesh
panchdrep
sompgome
flindun
fleepidge
wodsmid
drinbist

bremnate
shempabe
lindshlipe
hadgemest
premtrest
plintnench
flampsirk
slerthosh
ploudtomp
marpebarme
trepfreme
blembemurt
greembleem
gurffarnard
plemberb
nopgesh
blunfard
sabrund
nolbem
gortfabe
mibshrene
beskteb
poultibe
chemplenk
slibnawt
floskush
frintren
poskmerd
tregnasp
yongfobe
jushbime
prilbick
frennunk
firtmeeg
sepblosh
blengbim
maffbraim
plenshorte
slompresh
flimpnisk
fappimose
shustmoush
moograiste
prabefupe
nerlanch
slenkaslesh
depwub

barmdeg
mernsem
molshap
bompnaw
brelren
garmsen
herngoum
jornfrem
kirtlade
larnmard
mergsebe
nipetreb
thipsern
quendort
nelstrop
spordret
terpwafe
wertyape
brastparn
drepegane
framstrad
gramesoge
streadneb
yortslare
horpraste
lermtraist
marskplarn
plertdarpe
slarnprebe
tromfreest
flortbraime
chelftarg
lasknane
blambip
vandgid
sarnkie
trosthaib
reafloune
phaingak
fentcralk
prabrol
dwamphom
molklarch
sprewnofe
mieldral
larmpeb
brugrask

nilldorp
trablusp
tooptond
rewhab
bermnug
rellcond
chaldlisk
vad
mip
flosh
dit
dool
dask
dess
bep
mape
geg
fosht
blun
nace
jep
dap
gade
gax
tind
yat
bine
brote
koid
grod
nerge
sket
plote
stulpe
dran
neam
thod
blimb
derg
mese
lotch
ropp
gorples
trush
sloke
slar
lansh

yurge
reen
launt
nouve
possle
crope
flent
dype
hib
oast
weaf
komot
drupe
nend
fope
hing
ove
sheam
byue
gome
corf
loag
pice
noute
pler
bory
dight
slee
yock
stemp
yeant
wiss
temba
stam
haddle
foth
shurd
deide
beal
ost
mub
blace
flime
plet
rop
nace
yeant