Differential Effects of Cognitive Load on the Processing of Numerals and Standard Scalar Quantifiers

A thesis submitted by

Qingqing Wu

to

the Department of Linguistics

in partial fulfillment of the requirements

for the degree of Bachelor of Arts with Honors

Acknowledgements

This project of more than one year in the making could not have been possible without the guidance, expertise, and patience of Dr. Jesse Snedeker. I would like to thank her for leading me through this long, didactic process and for teaching me not only to ask the right questions, but also the best way to ask—as well as to answer—them.

I am also extremely grateful for the time, encouragement, and coaching provided by Manizeh Khan. She helped me navigate my way through the nebulous world of R, SPSS, Excel pivot tables, and taught me many other skills that I will take away from this thesis. I would like to thank the Harvard Lab for Developmental Studies for allowing me to use its resources and space to conduct my research.

Thank you to Dr. Maria Polinsky, Lauren Clemens, Greg Scontras, Andreea Nicolae, and my peers in the Department of Linguistics for cultivating my interest in semantics-pragmatics and offering insightful comments and suggestions throughout the thesis process.

Thank you to the Mary Gordon Roberts Fellowship and the Mind/Brain/Behaviour Department for funding my summer research on campus.

Finally, I would like to express my deepest gratitude to my friends and family for their continued support in all forms, whether it be going on coffee runs in my stead, listening to me think out-loud about an admittedly esoteric subject matter, or allowing me to take over their computers in times of need. Thank you.

Table of Contents

Abstract

Scalar quantifiers are often used as a case study to illustrate the interaction between semantics, the truth-conditional content of utterances, and pragmatics, the inferential analysis component of language. Both semantics and pragmatics contribute to the interpretation of standard scalar expressions such as "some", thereby allowing them to have two interpretations: the *weak* reading ("some and possibly all") and the *strong* reading ("some but not all"). Another class of scalar quantifiers, numerals, is often compared with standard scalar s because they, too, can have two interpretations: the *exact* reading, ("two" means "two and no more") or the *at-least* reading ("two" means "two and possible more"). A central debate in experimental linguistics has been on comparing the roles semantics and pragmatics play in the mechanisms of processing these scalar quantifiers. This thesis largely focuses on one area of processing: the speed and effortfulness of pragmatic inferences. To this end, I introduced a dual-task paradigm in which participants memorised letter sequences while doing a visual world eye-tracking task. The aim of this method was to disentangle the components of quantifier processing disrupted by working memory load; previous literature suggested that cognitive load would impair pragmatics, but leave semantics intact. My findings showed that the effect of cognitive load differs between standard scalars and numerals: for the former, the slightest load seemed to disrupt and delay the semantic analysis of "all" and "some", whereas for the latter, semantic analysis was robust and rapid across all load conditions. The results of this thesis could contribute to the ongoing discussion on the systematic processing differences between quantifier types.

Chapter One: Theoretical Background

1. Introduction to the semantics-pragmatics interface

Why and how do the different meanings of words arise? A prevalent, but inconspicuous phenomenon is that in certain instances, "some" could actually refer to "all", and "two" could also refer to "three". The difference between *language* and *communication* is an intriguing aspect of human interaction. In some cirumstances, we could strictly consider word meanings and the manner in which they combine to construct meaningful sentences: for example, in the following dialogue, the guest makes a direct response to the host, and no extra assumptions are necessary to interpret it.

(1) Host: What do you want to eat?

Guest: I want to eat cake.

However, there are often additional levels of meaning in everyday communication, some of which have become so ingrained in our conversations that we interpret them seemingly automatically. Notice the contrast between dialogues (1) and (2):

(2) Host: What do you want for dinner?

Guest: I like to eat cake.

On the surface, it seems that the guest's response does not map exactly onto the question posed by the host, as the guest's indication of her preference for cake does

not answer the host's inquiry about her dinner plans. However, there is an implicit component of the guest's utterance in response to the question posed indicating a desire to eat cake for dinner. So while the truth conditional content of the words in dialogue (1) is sufficient for its interpretation, the same could not be said for dialogue (2). For (2), we need an extra, context-dependent level of analysis—as basic as it may seem—to derive the implied meanings in the exchange.

In linguistics, we refer to the truth conditional content of words, which can be directly calculated from meanings of words and their structural relationships, as semantics. On the other hand, pragmatics denotes the more contextually-dependent aspect of communication derived via inferential analysis of speaker goals. The division between these two aspects of language was highlighted by Grice (1975). It is clear that these levels of representation interact in a close way. Thus, the difficulties lay with ascertaining where semantics ends and pragmatics begins. Furthermore, the diversity and prevalence of pragmatic inferences have made determining their mechanism of computation challenging.

A prominent operational difference between semantics and pragmatics is that pragmatic inferences are optional or cancellable, while truth conditional content is not (Sauerland, 2012). For example, in (2) above, the inference implying the guest's desire to have cake for dinner is optionally applied and can be defeasible in certain contexts (for example, if the guest followed his utterance in (2) with "but I'll save it for dessert").

2. Scalar implicatures

A case study at the semantics-pragmatics interface is the phenomenon of scalar implicatures, which are a type of pragmatic inferences made when interpreting expressions whose semantic informativeness varies on a scale (Horn, 1972). A type of these expressions is standard scalar quantifiers: <none, a few, some, many, most, all>. In particular, "some" often has two possible interpretations: the *strong* reading, in which it refers strictly to a proper subset of the total set ("some but not all"), or the *weak* reading, in which it is compatible with the total set ("some and possibly all") (Grice, 1989). Theorists have attempted to elucidate the underlying mechanism and the range of this phenomenon. Consider (3) and (4):

(3) Some apples are fruits.

(4) a. Some (and possibly all) apples are fruits.

$\exists x\ [apple(x) \cup fruit(x)]$

b. Not all apples are fruits.

$\neg \forall x\ [apple(x) \cup fruit(x)]$

(4a) delineates the meaning of (3) based on pure semantics, while (4b) incorporates the strengthening pragmatic inference. Though the reading in (4b) might be more salient in everyday conversation, the interpretation in (4a) could also be acceptable, because despite it being suboptimal—there is clearly an alternative quantifier available that is more apt to describe the situation—it is true that merely a subset of

apples could be referred to as fruits. Underinformative utterances such as (3) highlight the distinction between semantics and pragmatics (Bott & Noveck, 2004).

2.1 Are scalar implicatures cognitively effortful?

One of the main questions this thesis seeks to explore focuses on the pragmatic enrichment involved in scalar implicatures. In linguistic literature, several schools of theory, such as Relevance Theory, Pragmatic Theory, Lexical Theory, and Grammatical Theory, have debated about the nature of this step. Supporters of neo-Gricean and Lexical theories contend that pragmatic inferencing is automatic, whereas supporters of Relevance and Pragmatic theories argue that pragmatic inferencing is delayed and requires additional processing. Much of the focus for these theories is on ascertaining the positions at which scalar implicatures occur or how they arise. However, they do not directly address the notion of effortfulness, because automatic processes are not necessarily effortless and delayed processes are not necessarily effortful. Nevertheless, it is important to be aware of the theoretical divisions that linguistics have made about scalar implicatures in order to address the issues of cognitive effort more directly.

2.2 Pragmatic & Relevance Theories

Grice (1989) proposed that perception of speaker intention plays a crucial role in the derivation of scalar implicatures. His Cooperative Principle—that we should assume the speaker to be a collaborative and rational participant in conversation—is central to this idea. According to Pragmatic Theory, when hearing

a specific utterance, we seek to understand why the speaker would choose one possibility over an alternative one. Then, we can evaluate the possible alternative utterances. Finally, we would come up with an explanation for the speaker's decision. Consider (5):

(5) Some students are athletes.

In line with the Pragmatic Theory, we would assume that the speaker is cooperative by applying the relevant Gricean Maxims (Grice, 1989):

(6) Maxim of Quality: Do not say that for which you lack evidence

(7) Maxim of Quantity: Make your contribution as informative as is required

By assuming (6) and (7), we could conclude that in (5), the speaker lacks the evidence to make a stronger assertion—one with a quantifier higher on the scale, such as "all students are athletes"—such that the alternate, weaker quantifier on the scale, "some", is the most appropriate one to employ in the situation. In sum, the Pragmatic Theory contends that scalar implicatures arise because we undergo a reasoning process when hearing a particular utterance, which consists of considering the alternatives and applying the appropriate strength of interpretation (Sauerland, 2012). However, this theory makes no claims about the effortfulness of implicatures-making; despite the intuitive correlation between defaultness and cognitive effort, this link is neither established nor veritable.

In Relevance Theory, Sperber & Wilson (2004) propose that the decision to make an implicatures depends on whether this step is sufficiently relevant to

processing the utterance's meaning. Since the non-pragmatic interpretation of "some" might lead to a suitable reading of an utterance, the scalar implicature would likely be considered a non-necessary, extra-inferential step. Although intuitively, this theory seems to address the effortfulness issue, such that the extra step could map onto processing effort, some researchers have proposed that in contexts for which scalar implicatures are very salient, they could occur very quickly and effortlessly (Breheny et al., 2006).

2.3 Neo-Gricean/Lexical Theory

Levinson (2000), a main proponent of Lexical Theory, disagreed with the idea that scalar implicatures are a type of pragmatics and contended that executing the same reasoning steps each time a listener processes scalar terms like "some", as stipulated by the Pragmatic Theory, would be inefficient. Instead, it would seem more efficient for the product of this reasoning to be stored in the lexicon and easily accessible whenever it is needed. That is, implicatures could be built into the lexical meaning of a scalar term, such that "some" would denote "some but not all" by default (Sauerland, 2012). To account for the defeasibility of implicatures, Levinson raised the idea that these pragmatic inferences could be cancelled by the relevant contextual information, but that this would be a subsequent step in processing.

In the context of processing cost, it seems that this neo-Gricean account would encompass the idea that scalar implicatures are fast and effortless because they are built into the lexicon. However, some linguists argue that the act of summoning the scale of which "some" is a member, which occurs because the lower-

bound/*weak* meaning is retrieved before the negation of the stronger scale mate ("all") is applied, could be effortful (Levinson, 2000).

2.4 Grammatical Theory

Grammatical Theory attributes implicature computation to the listener applying a silent grammatical operator, *Exh*, which triggers an implicature at any level where it might be applied. Applying *Exh* to a proposition, *P*, would be defined as the conjunction of *P* and an epistemically *strong* implicature (Chierchia, 2004). The overt analogue of *Exh* is the operator "only". Grammatical Theory differs from Pragmatic Theory because it divorces implicature computation from reasoning about speaker intentions, and is rather a grammatical process. In this account, *Exh* is free to apply to embedded propositions, as in (9), whereas in Pragmatic Theory, *Exh* is applied to the entire sentence. Thus, in the Grammatical Theory, implicatures can be a part of the meaning of an embedded constituent.

(8) *Exh* (*Exh* (Mary ate the white chocolate or the dark chocolate) or she ate both chocolates)

In this view, computing a scalar implicature would involve 1) making the decision to apply *Exh*, and 2) deriving alternatives to the relevant scalar expression, and 3) excluding the non-applicable, weaker alternatives (Marty et al., in submission). Grammatical Theory focuses on the applicability of *Exh* to different components of an utterance rather than makes any predictions about the effortfulness of scalar implicatures calculation.

3. Are numerals processed the same way as standard scalars?

3.1 Similarities between standard scalars and numerals

Another common type of scalar expression is numerals, which exist on an intuitive and prevalent gradient (i.e., the number line) that most language users acquire from an early age. Interestingly, like standard scalars, they can also give rise to ambiguous readings. Consider (8) and (9):

(9) There are three ripe apples in the basket.

(10) a. There are three (and no more) ripe apples in the basket.

b. There are three (and possibly more) ripe apples in the basket.

(9) could be interpreted with an *exact* reading, as in (10a), or an *at-least* reading, as in (10b). There is a debate regarding whether numeral comprehension involves the same mechanism as that in scalar implicatures for standard scalars. Some linguists (e.g., Levinson, 2000) draw a parallel between deriving the *exact* reading for numerals and the *strong* reading for standard scalars: that (10b) would be the default interpretation for (9), and following the Gricean Maxims, the listener draws the inference that the speaker would have said "There are four ripe apples in the basket" if s/he believes that were the case, and only uttered (9) because s/he has no evidence for otherwise (Marty et al., in submission).

Numerals and standard scalars are also similar in the way their interpretation is affected by polarity (Panizza et al., 2009). Compare (11) and (12):

(11) a. Sara scored some/two points and she will win a prize.

b. Sara scored some/two points but not all/more and she will win a prize.

(12) a. If Sara scored some/two points, she will win a prize.

b. If Sara scored some/two points but not all/more, she will win a prize.

In downward-entailing contexts—linguistic environments that license inferences from a set to its proper subset—like those seen in (11), the lower-bounded readings of both scalar quantifier types seem to be more salient. Contrastingly, in upward-entailing context—linguistic environments that license inferences from sets to supersets—like those seen in (12), the upper bounded (i.e., *strong*) readings of both scalar quantifier types seem to be more salient (Panizza et al., 2012).

3.2 Differences between standard scalars and numerals

However, there are clear distinctions between these two types of quantifiers in other circumstances. For instance, numerals can bear an "exact" reading even when standard scalar items do not trigger *strong* readings, such as in negative and downward-entailing environments (e.g., conditionals, questions). Examine (13)-(16) (from Breheny, 2008):

(13) Do you have three children?

(14) a. No, I have two.

b. No, I have four.

?c. Yes. In fact, I have four.

(15) Do some of your friends have children?

(16) a. Yes. In fact, all of them do.

?b. No. In fact, all of them do.

In (13)-(14), the *exact* reading of "three" is more intuitive and pragmatically felicitous in this downward-entailing environment. However, in (15)-(16), the "some" does not garner a *strong* reading, and the *not-all* inference is not applied in the interpretation of the question in (15). This evidence counters the notion that numerals and standard scalars are processed similarly.

Why is there such incongruity between the two common scales? Some linguists theorise that although numerals and standard scalars have parallel processing mechanisms, numerals are on a more practised, natural, and ubiquitous scale than are standard scalars, such that alternatives on the scale are more easily accessible (Barner et al., 2011). This problem has been examined experimentally to certain extents from developmental and cognitive perspectives, to be discussed in Chapter Two. Thereafter, Chapter Three outlines the experimental design, Chapter Four and Five explains the results of the experiments, and Chapter Six is a discussion on the implications of the data on our understanding of these two types of quantifiers.

Chapter Two: Experimental Background

1. Connecting linguistic theory with experimental evidence

Researchers interested in understanding language meaning and the cognitive aspects of its representation have been attempting to reconcile linguistic theory and experimental evidence using methods from the discipline of psychology. It can be difficult to bridge the gap between the two areas of study, as different tools are used in their respective problem solving processes. Nevertheless, there is a growing body of experimental work on the semantics-pragmatics interface that has focused on determining the mechanism of scalar quantifier processing.

For standard scalars items like "some", experiments were initially designed to determine the extent to which semantic analysis and pragmatic inferences are involved in attaining the upper-bounded (*strong*) readings. A means of assessing this is through evaluating the effortfulness of implicature computation, as we work under the assumption that scalar implicatures are a type of pragmatic enrichment that incurs a processing cost (e.g., Marty et al., in submission). For bare numerals, psycholinguistics intended to ascertain whether the *exact* or the *at-least* meaning represent the default for number interpretation. Two main types of experiments have been conducted to this end: one examining developmental changes in the construal of these scalar items, and the other investigating the time course of scalar item interpretation in adults. This chapter evaluates several landmark studies that

have led the effortfulness debate in order to form a cogent basis for the experimental work proposed by this thesis.

2. Determining the speed and effortfulness of standard scalars

2.1 Making theories testable

Psycholinguists have proposed that there are separable stages involved in scalar implicature computation, thereby enabling experimenters to make testable hypotheses for the effortfulness of scalar implicature computation (adapted from Marty & Chemla, 2011). One framework for the stages is:

1. Semantic composition of the scalar item with its *weak* meaning occurs in context
2. Strengthening of the *weak* meaning of the scalar item occurs via implicature
3. If licensed by linguistic or extra-linguistic reasons, cancellation of the *strong* interpretation is carried out

If comprehension started with stage 1, such that strengthening of the scalar items to an upper-bounded meaning would not apply automatically, scalar implicatures computation would be delayed and effortful in most contexts. In contrast, the *weak* meaning of scalar items ("some and possibly all") could be accessed directly without going through stages 1 and 3, and would thus be a fast procedure. However, if stage 2 is automatically applied to the *weak* meaning of the scalar items, independent of context, processing would be fast and effortless, like lexical retrieval. In this case, attaining the weaker, literal meaning would require an

extra processing stage to cancel the "not-all" implicature, which would be demonstrable as a delay in processing.

It should be noted that not all theories from Chapter 1 agree with the existence of these stages, or the way by which they are related. For example, supporters of the Relevance Theory suggest that it is possible that both the *weak* and the *strong* interpretations of standard scalars like "some" are available from stage 1, but that in context, one of the readings become more *relevant* and is thus applied (Chierchia, 2004). However, for the purposes of focusing on the speed and effortfulness of scalar implicatures, we are going to assume that the above stages are applicable.

2.2 Experimental evidence for the effortfulness of scalar implicatures

Bott & Noveck (2004) tested these ideas in a study where participants completed truth-value judgment tasks involving underinformative sentences (e.g., "some elephant are mammals") and were evaluated for response time. The experimenters found that participants who made the pragmatic inference (and judged underinformative sentences such as the example from above as false) took longer than those who made the logical inference according to its semantic meaning. They attributed this difference to the time that it would take to generate the implicatures. This observation supports the idea that scalar implicatures are cognitively effortful, as the delay in reaction time can be accounted for by the application of the strengthened, upper-bounded meaning of the standard scalar.

However, other researchers have pointed out potential issues with judgment paradigms such as the one see in Bott & Noveck (2004). They frequently involve underinformative sentences, for which both logical (using the *weak* reading) and pragmatic (using the *strong* reading) judgments correspond to semantically valid interpretations of the quantifier. To link increases in reading time to one interpretation, the experimenters had to either manipulate the participants' understanding of the scalar quantifier or measure spontaneously occurring differences in the responding variable, both of which are suboptimal (Huang & Snedeker, 2009b). For the former, it would have been difficult to ascertain whether the processes involved in judging underinformative sentences out of context would be the same as those involved in ordinary comprehension. For the latter, the inference made would, at best, be a correlational one. This third-variable problem—in this case, the possibility that differences in reaction time between the two responses could be attributable to a mediating factor responsible both for the longer reaction times and for the contrasting responses—poses a challenge for this analysis (Huang & Snedeker, 2009b).

To circumvent these problems, Huang & Snedeker (2009a) used a procedure that could provide an indirect measure of comprehension while it takes place, the visual-world eye-tracking paradigm. This paradigm yields a sensitive, time-locked measure of linguistic processing. Participants were presented with spoken instructions asking them to manipulate objects within a visual scene, while their eye movements to those objects were measured. An advantage of this method is that eye movements are typically made without conscious reflection, which allows for a

more implicit measure of comprehension prior to—and perhaps distinct from—overt judgments, which in contrast may invoke higher-level strategic processes. Moreover, because eye movements are rapid, frequent and tightly linked to the processing of spoken language, they provide a fine-grained measure of how interpretation unfolds over time, such that they can provide information about the nature of comprehension at a given point (Huang & Snedeker, 2009a).

In Huang & Snedeker (2009b), participants were presented with stories auditorily and visually in which two types of objects were divided up between four characters, two boys and two girls. The items were always divided such that one of the critical characters (e.g., the girls) had a proper subset of one item (e.g., the socks) while the other had the total set of second item (e.g., the soccer balls). After the story, participants were given instructions like ‘‘Point to the girl that has some of the socks” and their eye movements were recorded. Notably, there was a period of semantic ambiguity beginning at the onset of the quantifier during which the referent of a lower-bounded reading of “some” was compatible with both of the critical characters. Eye movements to the target in this condition were compared to those in trials asking for ‘‘all of the socks” (in a context where one participant has all the socks and another has a proper subset of the soccer balls). In this case, the competitor character (the girl with some-but-not-all of the soccer balls) was inconsistent with the semantics of the quantifier. If the participant used the semantic content available to constrain interpretation prior to calculation of a pragmatic implicature, there would have been quick referential disambiguation in the “all” trials but prolonged competition between the two characters during the

"some" trials. To ensure that differences between these trials were not simply due to preferences for larger quantities or a greater difficulty in calculating upper-bounds relative to lower-bounds, Huang & Snedeker (2009a) also included conditions using "two" and "three". This was done under the assumption that numerals would not require a pragmatic inference to specify the *exact* meaning, and thus would not have the same temporary semantic ambiguity as "some". This allowed the "two" trials to act as a key comparison with the "some" trials. The experimenters found that eye movements to the referent targets were comparatively delayed to the upper-bound "some" compared to the quantifier without an implicature (i.e., "all") (Huang & Snedeker, 2009a).

These findings gave support to the theory that scalar implicatures require additional, pragmatics-related processing that is distinct from and occurs after semantic analysis. However, some researchers such as Grodner et al. (2010) suggested that delayed referential resolution for upper-bounded "some" compared to "all" in Huang & Snedeker (2009a) could be attributable to the salience of numbers as better quantity descriptors than the scalar items. Grodner et al. (2010) replicated this experiment without the numeral trials and found no relative delays in "some" compared to "all". Thus far, it is clear that the debate on the mechanism by which scalar implicatures occur is far from being settled, as there still seems to be several possible confounding variables that are preventing researchers from painting a clear picture of scalar implicature computation.

2.3 Contextual effects on scalar implicatures

Some psycholinguists predicted that the effortfulness of scalar implicatures is modulated by contextual effects: certain contexts could promote, and thus speed up, scalar implicatures, while others would not. Using a self-paced reading task, Breheny et al. (2006) examined the effects of context on the generation of implicatures. The advantages of this paradigm include greater temporal resolution and fewer demands on participants (Huang & Snedeker, 2009a). Participants were presented with the upper-bounded or lower-bounded context seen in (17) and (18) and their reading times were compared during two critical regions following the quantifier.

(17) Upper-bounded context: Mary asked John whether he intended to host all 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.

(18) Lower-bounded context: 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.

Participants in the upper-bounded context condition showed delays in reading the quantifier phrase (‘‘some of his relatives”), which suggests that the scalar implicature was calculated during this initial period. In contrast, participants in the lower-bounded context demonstrated delays in the region after the quantifier phrase, in which the proper subset was explicitly referred to (‘‘the rest would stay”),

which suggested that the upper-bounded inference had not yet been made in the initial period.

Though the results of this study seems to show that upper-bounded contexts can facilitate scalar implicature calculation immediately, some researchers (e.g., Huang & Snedeker, 2009a) have called the methodology of this experiment into question. For instance, the upper-bounded context not only highlights the presence of a contrasting quantifier in the context sentence (in (17), the stronger alternative on the scale, "all" is explicitly mentioned), but it also contains more overlap in information between the context sentence and the target sentence (there are more repeated words in (17) than (18)). This could have impacted reading times in the critical regions independent of effects on implicatures because the need to compute a scalar implicature was made more salient in the upper-bounded context sentences, independent of the manipulated variable of context.

Hartshorne & Snedeker (in submission) endeavoured to verify the results from Breheny et al. (2006) while eliminating the possible experimental confounds. To ensure that there were minimal differences between contrasting contexts, they used matched declarative and conditional sentences, the former being upward entailing and the latter being downward entailing. They found that although scalar implicatures were contextually dependent, this effect did not emerge immediately, and was in fact, still slow and effortful.

2.4 Developmental evidence for the effortfulness of scalar implicatures

Developmental studies of scalar implicatures provide a naturalistic way for psycholinguists to analyse the processing of scalar implicatures because studies have shown that children have trouble detecting ambiguity in referential communication tasks (Katsos & Bishop, 2011). This feature of language development, which has been attributed to a failure to employ the Gricean Maxim of Quantity, has manifested through their poor pragmatic competence, notably for scalar terms, such as modals (<might, must>) and conjunctions (<or, and>)(e.g., Noveck, 2001; Papafragou & Musolino, 2003). Because the process of pragmatic enrichment is more clearly separated from semantics in children than in adults, psycholinguists have posited that children's comprehension of scalar items would shed light on the effortfulness of scalar implicatures.

In experiments involving scalar implicatures, children tend to be literal in their interpretation of utterances and often fail to generate robust inferences (e.g., Noveck, 2001). For instance, Papafragou & Musolino (2003) found that five-year-olds, but not adults, were more likely to accept the usage of the weaker scalar, such as "some," in situations where the stronger term on the scale, such as "all," also applied (see also Barner et al., 2009; Noveck, 2001; Pouscoulous et al., 2007).

However, some researchers disagree with this analysis on the basis that the studies' methodologies have several limitations. For instance, most of these studies employed judgment tasks that required children to explicitly reason about another character's statement. In Papafragou & Musolino (2003), adults and children saw a

scene in which a girl finished a puzzle and was asked to evaluate whether the statement "The girl started the puzzle" was an apt description of the situation. Thus, it is possible that these tasks measured the participants' metalinguistic ability to reason about the felicity of using the weaker term rather than directly assessing whether children were making the pragmatic inference to express that "started" could mean "not finished" (e.g., Papafragou, 2006; Pouscoulous et al., 2007). Katsos & Bishop (2011) argued that five-year-old children are actually aware of underinformativeness, but are also tolerant of pragmatic infelicity, as they are less likely to judge underinformativeness as logical falsity. Though these experiments did not directly address the notion of cognitive effort in scalar implicature calculations, they corroborated the fact there is a substantial separation between pure semantics and the application of pragmatic inferences.

3. Determining the speed and effortfulness of processing numerals

3.1 Drawing parallels between numerals and standard scalars

To determine the means by which we comprehend numerals, psycholinguists have compared these scalar quantifiers to standard scalar items so as to test the idea that the *exact* meaning of numbers is attained in a mechanism similar to scalar implicatures. From a developmental perspective, if this was the case, then the widely documented difficulties that children face with making pragmatic inferences should also apply to making the exact reading for numerals. That is, we should expect that children would accept lower-bounded interpretations, even in contexts where adults prefer exact interpretations (Huang et al., in submission). Studies have

accumulated evidence contrary to this idea. Papafragou & Musolino (2003) found that while five-year-olds, but not adults, were content to accept weak scalar expressions (e.g., "started", "some") in situations where the stronger scalar term (e.g., "finished", "all") applied. In contrast, children, like adults, did not accept underinformativeness with numbers, and rejected sentences like "two of the horses jumped over the fence" in contexts in which they saw exactly three jump. The experimenters concluded that in contrast to how children comprehend other scalar items, they readily assign exact interpretations to number words. This conclusion is further supported by other comparable studies (e.g., Hurewitz et al., 2006; Pouscoulous et al., 2007).

Huang & Snedeker (2009) directly addressed the issue of the processing cost of numerals in adults by comparing them with standard scalars. They found that unlike "some", "two" and "three", were processed without delay in the on-line tasks, thereby suggesting that numerals carry an *exact* semantics. Panizza et al. (2009), using a similar paradigm, further differentiated standard scalars and numerals by determining that the stronger, exact meaning of "two" is accessed immediately in both upward entailing (UE) and downward entailing (DE) contexts, whereas the implicature for "some" is made more readily in UE contexts than in DE contexts. Therefore, it seems that while comprehending standard scalars is cognitively effortful, processing numerals is not.

Psycholinguists have proposed explanations for the difference between numerals and standard scalar terms. For instance, it could be possible that numerals

are more lexically focused than are standard scalars, which can yield more robust implicatures in the presence of contextual information focus (Bott et al., 2012; Zondervan, 2010). Furthermore, it is conceivable that the delay in scalar quantifier processing could be accounted for at least in part by the need to identify the alternatives on the semantic scale (< some < many < most < all), which might be more difficult to access than the numerals scale—knowledge that becomes ingrained in us from an early age and is likely acquired before other scales (Barner et al., 2011; Bott et al., 2012; Marty et al., in submission). However, the examination of the differences in the effortfulness of quantifier processing has thus far lacked direct evidence that can be analysed in real time. In Chapter Three, I propose a dual-task paradigm that could satisfy this condition.

Chapter Three: Experimental Design and Data Collection

1. Another method of comparing quantifier processing mechanisms

Much of the experimental data suggest that there is a relationship between ability to make pragmatic inferences and access to cognitive resources. Developmental studies, such as those conducted by Noveck (2001) and Huang & Snedeker (2009), show that children, for whom it is more difficult to apply pragmatic enrichment, tend to compute the *weak* reading for standard scalars and the *exact* reading for numerals. Latency studies, such as Bott & Noveck's (2004) experiment involving underinformative sentence judgments, suggest that imposing a time limit in the trials results in fewer pragmatically enriched responses, which could be explained by time restraints limiting the availability of resources allocated to implicature making. These findings indirectly lend support to the idea that making pragmatic inferences is associated with cognitive effort in quantifier processing.

1.1 Working with a dual-task paradigm

De Neys & Schaeken (2007) used a dual-task paradigm in order to test the role of working memory in computing scalar implicatures. They reasoned that by burdening subjects' cognitive resources with a demanding working memory task during the sentence comprehension task (similar to ones seen in Bott & Noveck, 2004), the default interpretation of the standard scalars would occur more frequently because the calculation of pragmatic inferences would be hindered by the

resulting reduction in cognitive resources. This is because the part of the executive control system that mediates analytic reasoning is cognitively demanding and draws on working memory resources. Notably. It usually overrides decisions made by a second automatic, heuristic system (De Neys & Schaeken, 2007). Theoretically, if working memory is loaded, the analytic system would be less able to control and overrule automatic, heuristic interpretations of the relevant quantifiers. In the experiment, De Neys & Schaeken (2007) first presented participants with a dot pattern that varied in complexity, asked participants to judge underinformative sentences while keeping the pattern in memory, and required them to reproduce the pattern thereafter. The experimenters found that the rate of scalar implicature computation decreased as cognitive load increased (i.e., as the dot pattern increased in complexity). These results further suggest that the pragmatic interpretation of standard scalars requires effortful, cognitive processing (De Neys & Schaeken, 2007).

Marty et al. (in submission) also applied the dual-task paradigm to compare the costs involved in processing standard scalars and numerals. Following the same logic as that of De Neys & Schaeken (2007), Marty et al. imposed a cognitive load on subjects who simultaneously comprehended numerals in context with the expectation that their pragmatic enrichment abilities would be hindered, such that the more default reading of numerals would arise. If the *exact* semantics account of numerals is accurate, then the retrieval of the meaning of numerals should not be affected significantly by the imposition of the cognitive load. However, if the *at-least* semantics account is valid, then a pragmatic step is required to arrive at an *exact*

interpretation of numerals in order to match the situation described in stimuli. Using the dual-task procedure for both standard scalars and numerals, Marty et al. (in submission) were able to determine whether arriving at the "strong" reading for standard scalars and the "exact" reading for numerals involved the same mechanism. Two modifications were made in this design compared to that of De Neys & Schaeken (2007): 1) instead of dot patterns, the cognitive load consisted of phonologically dissimilar letter sequences, the lengths of which could be manipulated to minimise or maximise the extent to which working memory would be occupied; 2) instead of underinformative sentences, subjects were asked to judge on a gradient the felicity of sentences that described the quantity of dots in pictures (most likely so that the context would be more natural for numerals compared to the alternative). Marty et al. (in submission) found that while the *strong* reading of standard scalars was less likely to be made under high cognitive load conditions, the same was true for the *at-least* reading of numerals, rather than the *exact* reading. They attributed this to the upper-bound of numerals being less effortful to derive than that of standard scalars because the former involves lexical retrieval, while the latter also necessitates pragmatic enrichment.

1.2 Implementing the dual-task procedure on-line

Nevertheless, the disadvantage of the sentence judgment tasks to evaluate quantifier processing, as discussed in Chapter Two, still exists with the aforementioned dual-task paradigm (in De Neys & Schaeken, 2007). In place of a binary evaluation of quantifier comprehension, I sought to use method that would allow a more fine-grained measure of the time course of comprehension via a visual-

world eye-tracking paradigm, as seen in the experiments carried out by Huang & Snedeker (2009). I aimed to directly map sentence processing onto separable periods of analysis to disentangle the aspects of interpretation that would be attributable to semantics or to pragmatics (Huang & Snedeker, 2009). Thus, while the dual-task procedure would allow us to deplete the cognitive resources needed to control the heuristic system—thereby revealing the automatic mode of comprehension—the visual-world eye-tracking paradigm would enable the comparison of the speeds and mechanisms by which standard scalars and numerals are processed.

2. Experimental design

I separated this study into two phases by the types of scalar quantifiers. In the eye-tracking task of each experiment, I evaluated the effect of Quantifier Strength, which contrasts the weaker ("some", "two") with its corresponding stronger ("all", "three") scalar term. In the cognitive load component of the dual-task paradigm, I used a letter sequence memory task, in which the length of the letter string is varied to establish three load conditions: no load, low (two-letter) load, and high (four-letter) load (see full list of letter sequences used in **Appendix A**). The load factor was manipulated between subjects.

2.1 Designing stimuli

The stimuli used in this study were based on those in Huang & Snedeker (2009) for the eye-tracking task and Marty et al. (in submission) for the cognitive load task. In the eye-tracking task, the backdrop featured two boys and two girls—

Mike, Julie, Phil, and Sarah—in four quadrants of the screen. They were arranged such that vertically adjacent characters matched in gender—that is, the boys were on the left side and the girls were on the right side. In each trial, each character would come into possession of two, three, or none of objects (see full list of objects in **Appendix B**). One of the horizontally adjacent boy-girl pair would receive a set of four objects divided equally amongst them, while one character of the other pair would receive a set of three objects of another kind and the other character of the pair would receive no object (refer to *Figure 1*).

Crucially, the two types of object in each object pair were designed to be compound nouns, such that they would share a period of phonological overlap for approximately two syllables at the words' beginnings. These items, such as "butterflies" and "buttercups", were contextualised in a relatable short story. Because of the compound nature of the nouns, we also had to ensure that the object pairs would not have set-subset relationships, such that one object could be identified as a superordinate set of the other object, because this ambiguity would be confusing for subjects in the eye-tracking task (e.g., we eliminated word pairs such as candy bar-candy-cane because the latter could also be labelled as the former). The prompt sentences included a relevant quantifier ("some", "all", "two", or "three") and one of the items in the word pair. The short stories and prompt sentences were recorded as sound files, and were programmed to accompany the visual stimuli (see example in *Figure 1* below).

- Short story: “The boys and girls went to the forest for a school fieldtrip. Julie and Mike saw lots of butterflies. Sarah saw lots of buttercups, but Phil didn’t see any.”
- Prompt sentence: “Point to the girl that has some/all/two/three of the butterflies/buttercups.”

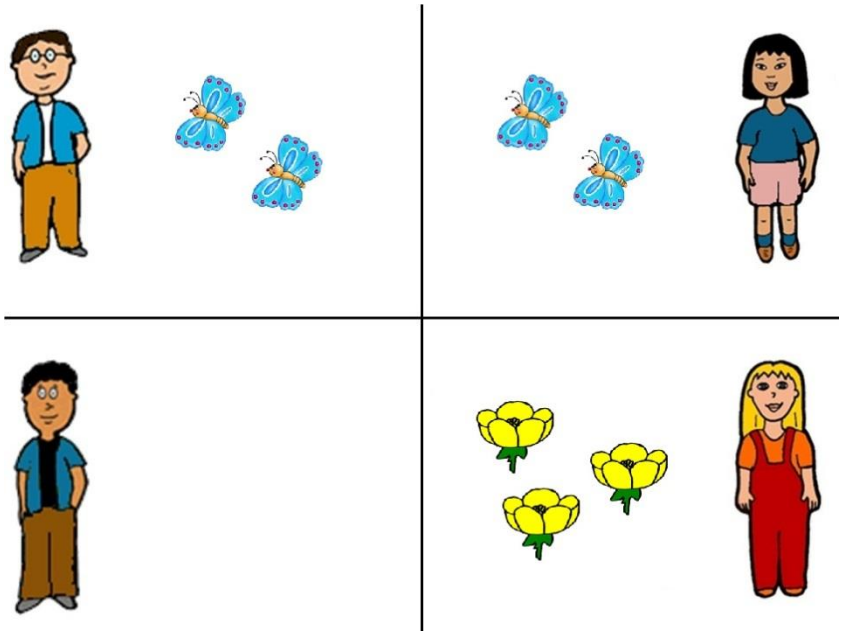

Figure 1. The visual display of a sample trial

The memory task is similar to the one introduced by Marty et al. (in submission). Ten phonologically dissimilar letters of the alphabet are chosen and arranged into random permutations of two and four string sequences. Sixteen distinct sequences are used for each load condition.

3 Experimental hypothesis

3.1 Standard scalars in the dual-task paradigm

Combining the cognitive load and visual-world eye-tracking paradigms could give us information on how semantic and pragmatic processes interact in real-time, and to ascertain whether an effortful pragmatic inference is made for standard scalars. Under the assumption that the heavier the cognitive load, the more automatic and heuristic processing would be, we could infer that pragmatic inferencing would be disrupted with the higher load trials. If scalar implicatures were not effortful, load should not create an effect of Quantifier Strength or Type, since load should not disrupt lexical retrieval because it does not entail pragmatic enrichment. This would manifest by a rapid increase of fixations to the target character relative to the competing character of the same gender as soon as the quantifier information is given. On the other hand, if scalar implicatures were indeed effortful, we would predict that the higher the load, the more delayed the scalar implicatures would be for "some" compared to low- or no-load trials. This is based on the notion that if quantifier processing is composed of semantic analysis before scalar implicature making, then we would expect fixations to be relatively equal between target and competitor characters when quantifier information is initially given, because at this point, both "all" and "some" could logically refer to the total set. Instead, we would expect disambiguation for the referent of "some" to be delayed until after the quantifier is presented.

3.2 Numerals in the dual-task paradigm

According to the *exact* semantics account of numerals, we should not see a significant difference between the processing of "two" and "three", since no pragmatic processing is required such that the load manipulation would be disruptive. Furthermore, the processing of numerals would be immediate. The *at-least* semantics account of numerals would predict a delay in numeral processing with the load manipulation because the pragmatics required to compute an *exact* reading would be cognitively effortful. This theory also predicts that to attain the *exact* interpretation of "two", a pragmatic inference would have to take place, which would manifest as a delay in processing compared to "three" during the high load trials.

In the context of a comparison of numerals against standard scalars, we expect there to be a significant difference between the two types of quantifiers if the *exact* semantics account of numerals were true. That is, according to this view, numeral processing should be relatively faster than that of standard scalars, since quantifier information should be incorporated into comprehension immediately, without pragmatic enrichment or extensive semantic analysis like standard scalar processing might require.

Chapter Four: Experiment 1

1. Experimental set-up

1.1 Norming tasks

Two norming tasks were created to select the stimuli used in the eye-tracking task. First, to ascertain whether the quantifier is predictive off-line, 16 native-English speaking participants completed a questionnaire in which they were provided with the shorts stories in the format outlined above, but with the label of the target object missing. They were also provided with an image of the quadrants in which Mike, Julie, Phil, and Sarah were pictured beside the items with which they have been distributed (as in *Figure 1*). Participants were asked to fill in the blank with the identity of the item matching the gender and the quantity given (e.g., "Point to the girl that has some/all/two/three of the _____"). Participant responses were taken into account when determining the list of word pairs to be utilised in the study. Out of the word pairs that passed the norming requirement, the final 16 word pairs were chosen based on the length of their phonologically ambiguous regions—the word pairs with the longest ambiguous regions were desirable to as to maximise the amount of time during which we might see anticipatory looks to target before it is explicitly disambiguated during the prompt sentence.

The second norming task was conducted with Amazon Mechanical Turk with the goal evaluate the felicity of the labels matched to the visual stimuli. Participants were registered workers on the site who self-identified as native English speakers.

For the task, they were presented with the same visual display such as the one above, where each of the four characters with a quantity of items is a possible choice to answer a given prompt (e.g., "Point to the girl that has some/all/two/three of the butterflies"). Responses were assessed for accuracy.

1.2 Counterbalancing stimuli lists

Both of the experiments involved the dual-task paradigm in which participants memorised letters while following auditory commands that involved quantifiers (see *3.2.1* for example of prompt sentence). Overall, the study was set up in a 3 × 2× 2 design, in which Quantifier Type and Load were manipulated between-subjects, and Quantifier Strength within-subjects. Four versions of the stimuli were created to counterbalance the quantifiers and target objects, such that only one version of each item would be seen by each subject. Each list contained three practice trials and sixteen test trials. Furthermore, each list occurred in three load conditions: no load, low load, or high load.

2. Methods

2.1 Participants

36 native English speakers between the ages of 18-25 who were recruited via Harvard University's Study Pool participated in this study. They received either course credit or $5 for their participation.

2.2 Procedure

Participants sat in front of a TOBII T60 eye-tracker and a keyboard. Following calibration, they were given detailed written instructions on the

computer screen, which made explicit the importance of reproducing the letter sequence correctly in the experiment. They were also instructed to listen carefully to the short story presented during the eye-tracking task of each trial and to follow the prompts in order to answer the pertinent question.

Each trial started with the presentation of the letter sequence to be memorised. Then, the quadrant display (such as that in Figure 1) would appear, with a pre-recorded female voice describing which items each character is to receive. Participants were asked to touch the character that matched the description given by the prompt sentence, which always included the gender of the target character and a quantifier noun phrase referring to the character's possessions. At the end of each trial, participants were asked to reproduce the letter string from the beginning of the trial in reverse order by typing it on the keyboard. Thereafter, they received feedback on the accuracy of their response. The general structure of a two-load trial is depicted in *Figure 2*.

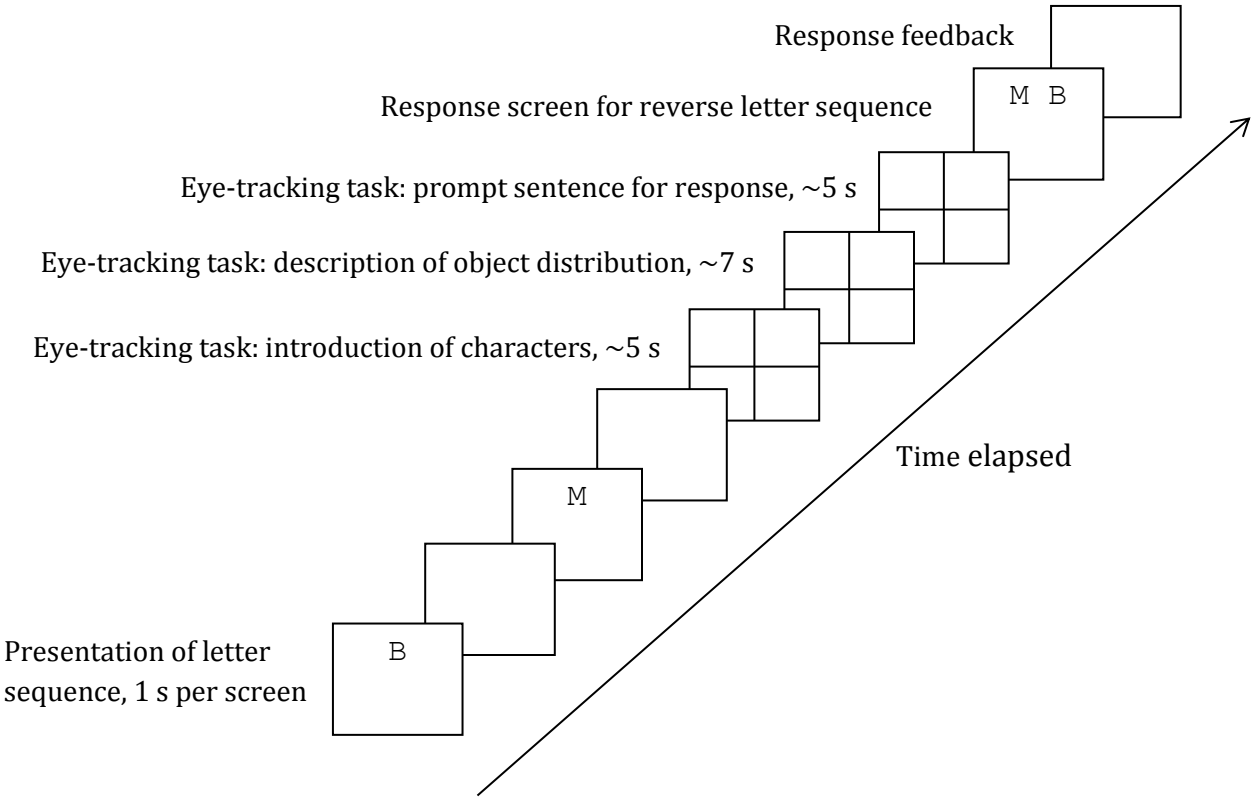

Figure 2. Sequence of events for one trial in the two-load condition

We made a notable modification to the procedure in order to appropriate implement the zero-load condition trials: the zero-load trials are identical to the two-load trials, except that the response screen for the recall of the letter sequence and the feedback screen for this task are placed immediately after the presentation of the letters and before the eye-tracking task. This is to ensure that the dual-task nature of the study is kept constant, but also to remove the cognitive load because participants would not be required to hold a letter sequence in their memory during the eye-tracking task.

2.3 Results

LOAD TASK

We analysed performance on the load task by assessing the percentage of letter that were retained (percentage-retention) of each trial. For the zero-load trials in Experiment 1, the mean percentage-retention was 100%, while it was 96% for the two-load trials and 98% for the four-load trials. There was no significant difference between the conditions ($F(1, 71) = 0.61$, $p > 0.46$). The high percentage-retention of the memory load indicates that during the eye-tracking task, participants' cognitive resources were used to the extent that our experimental manipulation demanded, thereby verifying that this component of the dual-task was applied effectively.

EYE-TRACKING TASK

During the prompt sentence of the short story, participants' gaze was tracked. We examined the proportion of participants' gaze to the target character and tracked the data over five sentence regions:

1. Sentence Start: This marks the onset of the prompt sentence and includes the words "Point to the".
2. Gender: This marks the period of the sentence that begins with the gender of the target character and ends with before the introduction of the quantifier. We expect participants' eye gaze to shift to the side of the screen with only the characters that matched the gender cue.

3. Quantifier: This region begins at the onset of the quantifier and ends before the target noun. For Experiment 1, the relevant quantifier is “some” or “all”. If the participants rapidly compute the scalar implicature, this is the earliest sentence region in which they should be able to identify the target character.
4. Noun Start: This region marks the introduction of the target noun. At this point, the referent is still ambiguous because there is a two syllable overlap between the target and competitor nouns during which we would be able to determine whether applying pragmatics occurs after semantics analysis. If so, we would still expect fixations to be relatively equal between target and competitor characters as pragmatic calculation would be done during this period.
5. Disambiguation: At the onset of this region, we are introduced to the part of the target noun that would allow clear distinction between the target and competitor nouns. In this period, eye gaze should shift to the target character regardless of scalar quantifier processing.

We analysed looking time to the target character as a proportion of total looking time to the target and competitor characters, with the competitor being the character that is of the same gender but a different object as the target. We focused our analysis on the three regions of the prompt sentence: Quantifier, Noun Start, and Disambiguation. We conducted t-tests and ANOVAs in order to determine 1) the sentence region during which participants started looking at the target character

above chance, 2) whether this depended on the factor of quantifier strength ("some" vs. "all"), and 3) whether cognitive load affected this.

Figure 3, which graphs proportion of looks to the target character collapsed across Quantifier Strength over the time course of the prompt sentence, shows that participants seemed to be delayed in using quantifier information from the standard scalars to resolve target/competitor noun ambiguities, as one-sample t-tests show that looks to target did not significantly exceed chance until the Noun Start region (see *Table 1*). However, this also shows that participants did use quantifier information to disambiguate references, since looks to target were significantly above chance before the Disambiguation region.

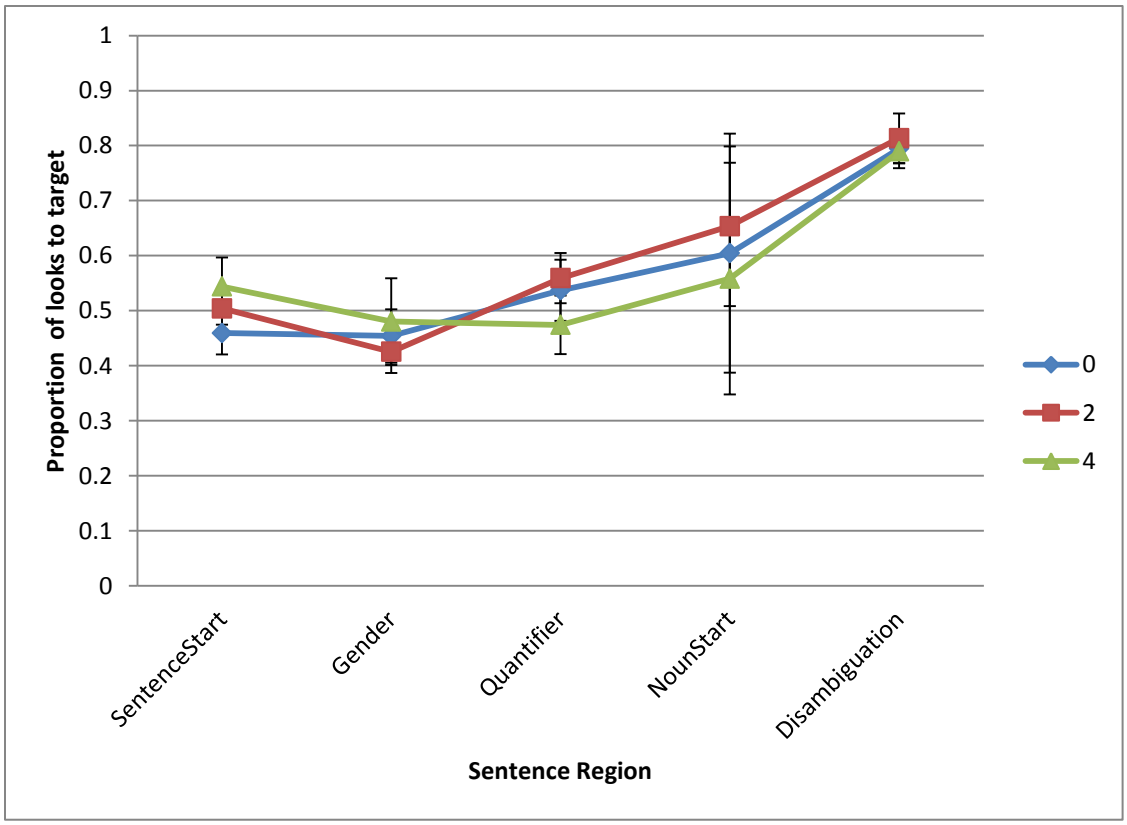

Figure 3. Experiment 1: proportion of looks to target during the prompt sentence by Load

Notably, according to *Table 1*, there seems to be a perceptual bias towards the subset: according to *Table 1,* looks to target are already nearly significantly above chance for the "some" trials at the Quantifier region ($t(35) = 1.98$, $p < 0.06$).

		Sentence Region		
Standard Scalar	**Measurement**	***Quantifier***	***Noun Start***	***Disambiguation***
all	proportion of looks to target	0.50	0.58	0.81
	t-test against chance ($t = 0.5$)	$t(35) = 0.10$ $p = 0.92$	$t(35) = 2.37$ $p =$ **0.02**	$t(35) = 14.90$ $p = 0.00$
some	proportion of looks to target	0.55	0.60	0.80
	t-test against chance ($t = 0.5$)	$t(35) = 1.98$ $p = 0.06$	$t(35) = 3.49$ $p =$ **0.00**	$t(35) = 9.34$ $p = 0.00$

Table 1. Experiment 1: proportion of looks to target in the critical sentence regions tested against chance

To further explore the interaction between Load and Quantifier Strength, we conducted repeated measure ANOVAs for the three sentence regions, with Quantifier Strength as a within-subjects variable and Load as a between-subjects variable (see *Table 2* and *Figure 4*). As shown in *Table 2,* Load and Quantifier Strength had no significant effects on looks to target at any of the critical sentence regions. However, there was a near-significant interaction between Quantifier Strength and Load starting at the Noun Start region ($F(2, 33) = 3.04$, $p < 0.06$) and continuing into Disambiguation ($F(2, 33) = 2.81$, $p < 0.07$).

	Sentence Region		
Measurement	***Quantifier***	***Noun Start***	***Disambiguation***
effect of Quantifier Strength	$F(1, 33) = 1.45$ $p = 0.24$	$F(1, 33) = 0.47$ $p = 0.50$	$F(1, 33) = 0.02$ $p = 0.89$
effect of Load	$F(2, 33) = 1.96$ $p = 0.16$	$F(2, 33) = 1.77$ $p = 0.19$	$F(2, 33) = 0.07$ $p = 0.93$
Quantifier Strength × Load interaction	$F(2, 33) = 1.58$ $p = 0.22$	$F(2, 33) = 3.04$ $p = \mathbf{0.06}$	$F(2, 33) = 2.81$ $p = \mathbf{0.07}$

Table 2. Experiment 1: repeated measures ANOVAs of proportion of looks to target in the critical sentence regions

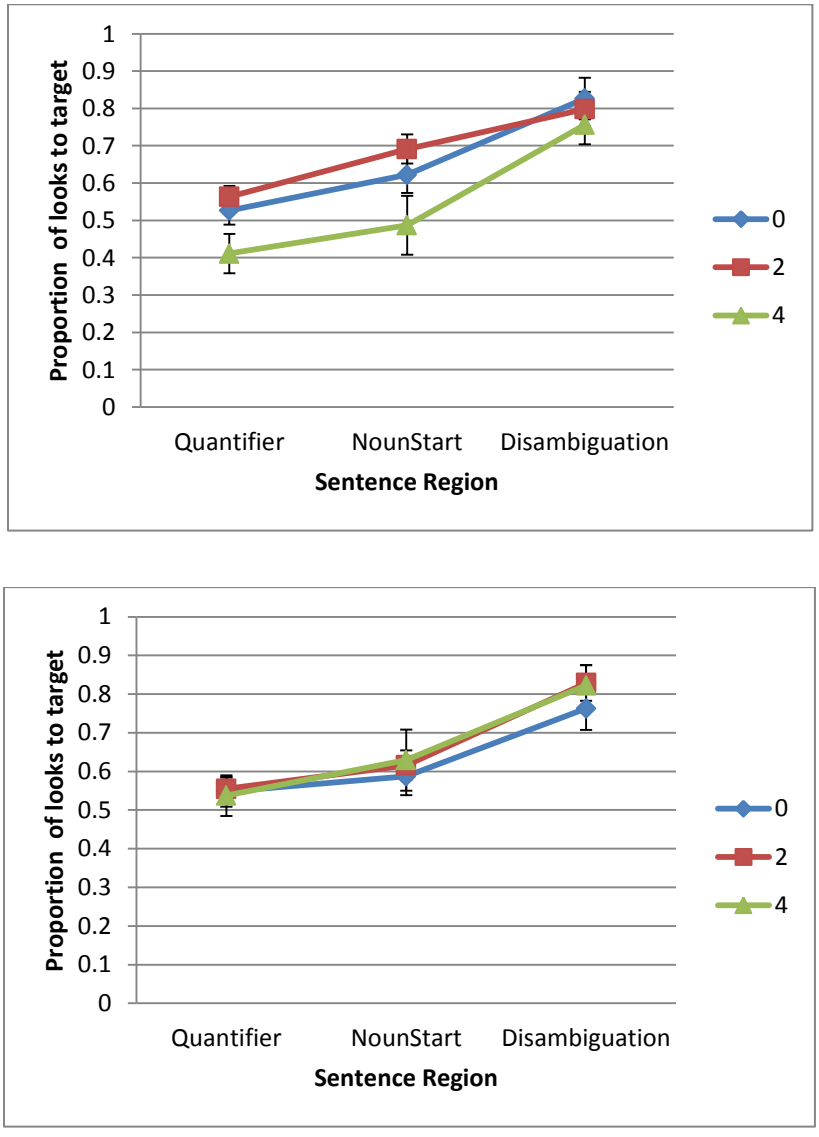

Figure 4. Experiment 1: comparing the effect of Load on "all" (top) and "some" (bottom) trials

To explore the basis of this interaction, we conducted separate ANOVAs to explore the effect of Load on the "some" and "all" trials. As seen in *Table 3,* Load had an effect on the "all" trials at the Quantifier and Noun Start regions, but not on the "some" trials at any sentence region.

	Sentence Region		
Measurement	***Quantifier***	***Noun Start***	***Disambiguation***
effect of Load on "all"	$F(2, 33) = 2.96$ $p =$ **0.07**	$F(2, 33) = 4.65$ $p =$ **0.02**	$F(2, 33) = 1.71$ $p = 0.19$
effect of Load on "some"	$F(2, 33) = 0.05$ $p = 0.95$	$F(2, 33) = 0.16$ $p = 0.85$	$F(2, 33) = 0.38$ $p = 0.70$

Table 3. Experiment 1: one-way ANOVA by Load of proportion of looks to target in the critical sentence regions

This effect can be clarified by zooming in on the Noun Start region: for the "all" condition, reliable referent resolution occurred by this region for the zero- ($t(11) = 2.06$, $p < 0.06$) and two-load ($t(11) = 4.18$, $p < 0.01$) trials, but not for the four-load trials ($t(11) = -0.87$, $p > 0.40$).

3. Discussion of Experiment 1

In Experiment 1, we found that although participants are able to use quantifier information to resolve ambiguous references under cognitive load. The incorporation of quantifier information was not immediate, as looks to target exceeded chance at the Noun Start region of the prompt sentence rather than the Quantifier region. However, there were several findings are hard to reconcile with previous literature. First of all, the cognitive load manipulation seemed to affect standard scalar processing to a greater degree than it has in comparable studies, the

most notable being Huang & Snedeker (2009). Whereas the eye-tracking task from Huang & Snedeker (2009) showed that "all" is disambiguated during the Quantifier region, thereby implying that participants were able to quickly use the quantifier information for reference resolution, the results from this experiment showed that the processing of "all" is delayed to a large extent, even in the low load conditions. Therefore, instead of solely impairing the processes involved in deriving scalar implicatures—which would have only impacted the interpretation of "some"—there was a clearly observed effect on "all". These results, which ran counter to the hypothesis, could be explained in a few ways. First, it is conceivable that whereas Load has a graded effect on the semantic analysis required to comprehend "all", this manipulation completely knocked out any pragmatic inferencing needed to calculate implicatures for "some". However, an interesting phenomenon was that the "some" trials did not seem to be as affected by this semantic disruption. This could be accounted by for a baseline subset preference (i.e., non-linguistic preference for the characters with two objects) that participants seemed to have, as seen in the fact that looks to target in the "some" trials were already close to above chance by the Quantifier region. A possible explanation is the lack of variability in the distribution of the objects to the characters: since the answer to the prompt sentence was never the character with no objects, the target was more likely to be situated on the half of the screen with the characters possessing two objects each, which corresponds to the quadrant with the character corresponding to "some". It is, however, difficult to explain why there was a discrepancy between the subset preference in this

experiment and the total set preference (i.e., non-linguistic preference for the character with three objects) seen in Huang & Snedeker (2009).

Another potential issue is that the cognitive load manipulation might have not been sufficiently sensitive to differentiate between semantic and pragmatic processing broadly. Compared to the dual-task paradigm from Marty et al. (in submission), which involved truth-value judgments as an offline linguistic task that required implicit semantic output before completing, the eye-tracking task in this experiment used an online measure, which involved the rapid prediction of the correct referent that eventually led to disambiguation. Therefore, there could have been a smaller window of opportunity for re-analysis and correction by executive control mechanisms. This was reflected in the difficulty this paradigm had with distinguishing between the extent to which semantics and pragmatics contribute to the interpretation of "some" and "all", since both processes were impaired, contrary to predictions we could make based on prior literature.

Chapter Five: Experiment 2

1. Goal of Experiment 2

Our findings from Experiment 1 demonstrate that the cognitive load component of the dual-task did not solely impair pragmatic processes. Instead, it seemed to also disrupt semantic processes, even in low-load conditions. These results raise new questions about why semantic processes (or the mapping of semantics onto the stimuli in the eye-tracking task) were disrupted by the cognitive load task. To address this question, we further explored the scope of this semantic disruption to ascertain if it is limited to standard scalar or if it would also influence the interpretation of numerals. In order to directly compare these two types of scalar quantifiers, we designed Experiment 2 to be minimally different from Experiment 1 by simply replacing the standard scalar with a numeral in the prompt sentences of the eye-tracking tasks.

2. Methods

2.1 Participants

36 native English speakers between the ages of 18-25 who were recruited via Harvard University's Study Pool participated in this study. They received either course credit or $5 for their participation.

2.2 Procedure

The procedure was identical to Experiment 1. A key difference between the experiments, however, was that the quantifier type tested was numerals ("two" and "three"), instead of standard scalars: "two" replaced "some" and "three" replaced "all" in the prompt sentences.

2.3 Results

LOAD TASK

Like in Experiment 1, we evaluated the load task by the percentage-retention of the letters per trial. For the zero-load trials in Experiment 2, the mean percentage-retention was 100%, while it was 97% for the two-load trials and 98% for the four-load trials. There was no significant difference between the two conditions ($F(1, 46) = 0.61$, $p > 0.46$). In addition, there seems to be no significant difference in the quality of the load task completion between Experiments 1 and 2 ($F(1, 46) = 0.47$, $p > 0.50$).

EYE-TRACKING TASK

The same analyses that were conducted in Experiment 1 was done for Experiment 2, focusing on the looking time to the target character as a proportion of total looking time to the target and competitor characters to evaluate participants' ability to resolve referential ambiguity during the prompt sentence. The sentence regions Quantifier, Noun Start, and Disambiguation were analysed. We conducted t-

tests and ANOVAs in order to determine how Quantifier Strength and Load affect numeral processing.

As shown in *Figure 5* and *Table 4*, looks to target reached above chance at Quantifier region for the numerals, thereby indicating that the lexical information from the numerals was used rapidly.

		Sentence Region		
Numerals	**Measurement**	***Quantifier***	***Noun Start***	***Disambiguation***
three	proportion of looks to target	0.63	0.77	0.91
	t-test against chance (t = 0.5)	$t(35) = 3.95$ $p = 0.00$	$t(35) = 7.82$ $p = 0.00$	$t(35) = 18.41$ $p = 0.00$
two	proportion of looks to target	0.66	0.85	0.93
	t-test against chance (t = 0.5)	$t(35) = 6.10$ $p = 0.00$	$t(35) = 15.18$ $p = 0.00$	$t(35) = 26.12$ $p = 0.00$

Table 4. Experiment 2: proportion of looks to target in the critical sentence regions tested against chance

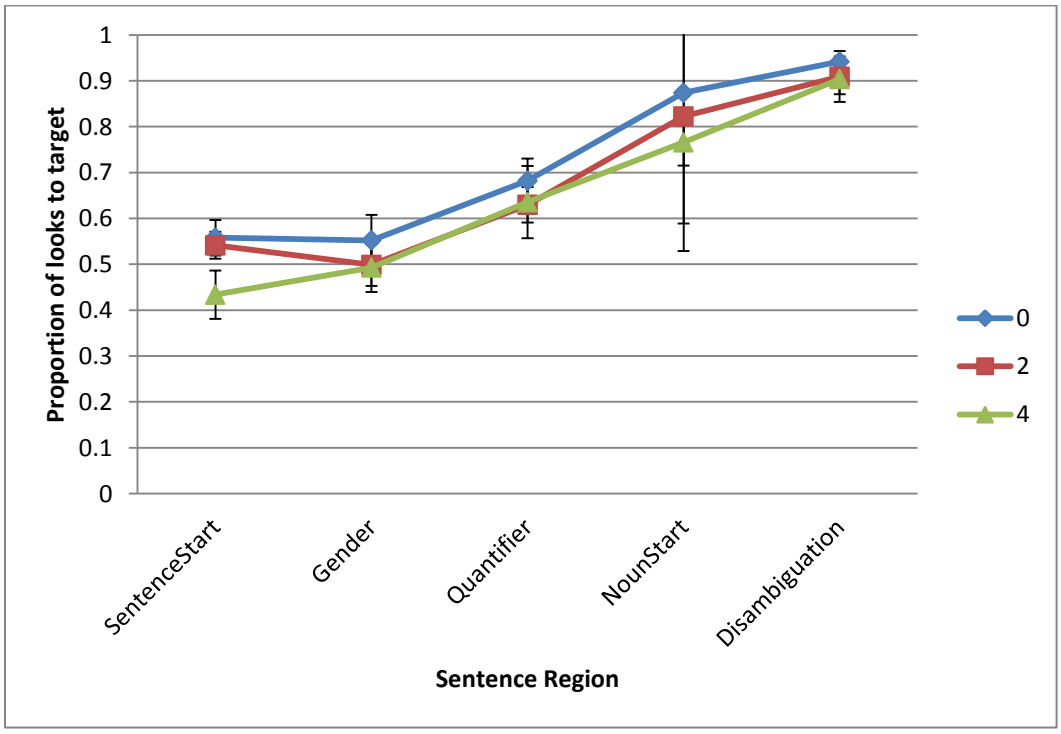

Figure 5. Experiment 2: looks to target during the prompt sentence by Load

Table 5 shows that there was no consistent effect of Quantifier Strength or Load, as indicated by the results from repeated measure ANOVAs for the three sentence regions. There was also no Strength by Load interaction at any sentence region. However, there is an anomalously significant effect of Quantifier Strength at the Noun Start region ($F(1, 33) > 4.89$, $p < 0.03$), which was not observed at any other critical sentence region. Closer examination of this phenomenon suggested that we may ascribe this to the same perceptual bias seen in Experiment 1 (for a comparison, refer back to *Table 4*).

	Sentence Region		
Measurement	***Quantifier***	***Noun Start***	***Disambiguation***
effect of Quantifier Strength	$F(1, 33) = 0.55$ $p = 0.46$	$F(1, 33) = 4.89$ $p =$ **0.03**	$F(1, 33) = 6.47$ $p = 0.43$
effect of Load	$F(2, 33) = 0.75$ $p = 0.48$	$F(2, 33) = 2.01$ $p = 0.15$	$F(2, 33) = 0.24$ $p = 0.79$
Quantifier Strength × Load interaction	$F(2, 33) = 0.59$ $p = 0.56$	$F(2, 33) = 0.21$ $p = 0.81$	$F(2, 33) = 0.88$ $p = 0.43$

Table 5. Experiment 2: repeated measures ANOVAs of proportion of looks to target in the critical sentence regions

3. Discussion of Experiment 2

As predicted by the *exact* semantics account of numerals, both "two" and "three" were disambiguated quickly across all load conditions, which indicates that the cognitive load task did not impact any aspect of quantifier processing, including the upper-bounded meanings of numerals. However, the subset preference—a higher chance that participants look to the characters with two objects for non-linguistic reasons—from Experiment 1 persisted, which could explain the effect of Quantifier Strength observed in the Noun Start region. In general, the results from

Experiment 2 regarding numeral processing are congruent with those from prior literature, which also found the *exact* interpretation of numerals to be immediate and undifferentiated between "two" and "three" (e.g., Huang & Snedeker, 2009).

4. Comparing Experiments 1 and 2

By directly comparing the data from both experiments, we can determine whether the mechanisms by which standard scalars and numerals are processed differ. *Table 6* presents the results of the 3×2×2 ANOVA that captures the possible effects between Experiments 1 and 2.

	Sentence Region		
Measurement	***Quantifier***	***Noun Start***	***Disambiguation***
main effect of Quantifier Strength	$F(1, 66) = 1.82$ $p = 0.18$	$F(1, 66) = 3.73$ $p =$ **0.06**	$F(1, 66) = 0.21$ $p = 0.65$
main effect of Load	$F(2, 66) = 1.28$ $p = 0.29$	$F(2, 66) = 2.95$ $p =$ **0.06**	$F(2, 66) = 0.42$ $p = 0.66$
main effect of Quantifier Type	$F(1, 66) = 16.60$ $p =$ **0.00**	$F(1, 66) = 39.99$ $p =$ **0.00**	$F(1, 66) = 15.70$ $p =$ **0.00**
Quantifier Strength × Load interaction	$F(2, 66) = 0.14$ $p = 0.87$	$F(2, 66) = 1.15$ $p = 0.32$	$F(2, 66) = 2.02$ $p = 0.14$
Quantifier Strength × Type interaction	$F(1, 66) = 0.05$ $p = 0.82$	$F(1, 66) = 0.75$ $p = 0.39$	$F(1, 66) = 0.43$ $p = 0.52$
Quantifier Type × Load interaction	$F(2, 66) = 1.37$ $p = 0.26$	$F(2, 66) = 0.82$ $p = 0.45$	$F(2, 66) = 0.16$ $p = 0.85$
Quantifier Strength × Type × Load interaction	$F(2, 66) = 1.89$ $p = 0.16$	$F(2, 66) = 2.66$ $p =$ **0.08**	$F(2, 66) = 1.14$ $p = 0.33$

Table 6. Comparing Experiments 1 and 2: repeated measures ANOVAs of proportion of looks to target in the critical sentence regions

The most remarkable effect observed is that looks to target for the two types of quantifiers differ significantly at all three critical sentence regions (all F's$(1, 66) > 15.70$, all p's < 0.00). Whereas participants in the numerals trials were able to use quantifier information to resolve referential ambiguity at the Quantifier region,

those in the standard scalar trials had more delay in this endeavour and did not disambiguate between competitor and target characters until the Noun Start region (see *Figure 7*).

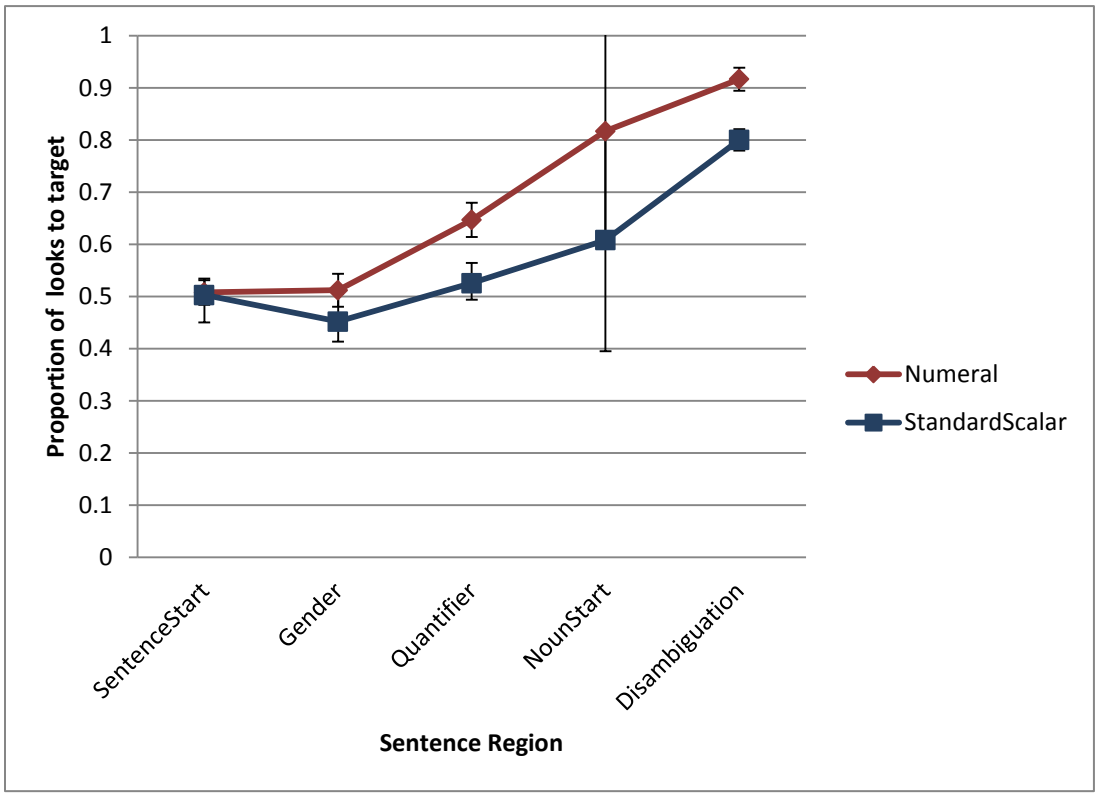

Figure 7. Comparing Experiments 1 and 2: looks to target during the prompt sentence by Quantifier Type

In addition to the main effect of Quantifier Type, there was also a near-significant effect of Quantifier Strength at the Noun Start region ($F(1, 66) = 3.73$, $p < 0.06$). This corroborated the pattern seen in separate analyses of the experiments: that there was a perceptual bias towards the subset (for the characters

corresponding to having “some” or “two” of the objects) during the eye-tracking task.

Interestingly, there is a near-significant Quantifier Strength × Load × Quantifier Type interaction ($F(2, 66) = 2.66$, $p < 0.08$). This piece of data, combined with the near-significant effect of Load ($F(2, 66) = 2.95$, $p < 0.06$), could be construed as Load having a different effect depending on the combination of the other two factors, Quantifier Strength and Quantifier Type. Furthermore, while Load significantly impacted the processing of standard scalars by impairing the semantic analysis of “all” (as seen in Experiment 1), Load did not have a significant effect on the processing of numerals. Therefore, this overall effect of Load appears to be driven by its impact on standard scalar processing.

Chapter Six: Discussion and Conclusion

This thesis began with the premise that a cognitive load would disrupt pragmatic processing without affecting semantic processing. Specifically, we predicted that if attaining the upper-bounded (*strong*) readings of scalar quantifiers would require pragmatic inferencing, their comprehension would be rendered effortful and delayed. However, we found that even the semantic processing of the quantifiers, in addition to pragmatic processing, was impaired by the load. From the eye-tracking data in Experiment 1, it appears that both standard scalars "all" and "some" were impacted by the cognitive load, but this effect manifested in different ways. The ANOVAs of the data seems to tell the story that the effect of Load was more significant for "all" than for "some". However, this is not an accurate conclusion. If we compare the time course of referential resolution during the eye-tracking task in Experiment 1 to that seen in a similar task in Grodner et al. (2010), we find that for both "some" and "all" trials in this study, the proportion of looks to target increased on a relative delay across both low- and high-load conditions. For "some", there was only a 5% increase (55% to 60%) in looks to target from the start of Quantifier to the end of the ambiguous Noun Start region, compared to a 8% increase for "all" (50% to 58%) (see *Table 1* and *Figure 4*). Notably, in contrast, Grodner et al. (2010) found an almost 20% (from 47% to 67%) increase in the corresponding time window for "some". This could be accounted for by the idea that the derivation of the upper-bound is impaired incrementally by the cognitive load for "all", the same

process is impaired almost completely for “some”. We were able to detect this effect for “all” because it was differentiable by the load condition.

Furthermore, a baseline preference that participants seemed to have for looking at the characters with the fewer (two) objects on the screen was detected in both Experiments 1 and 2. This is perhaps explicable by the fact that since the prompt sentence never asks for the characters with none of the objects—which coincides with the half of the screen where the character with the total set (“all” or “three”) of the objects was situated—the answer to the prompt sentence is more likely to be the characters with the subset (“some” or “two”).

Why and to what extent is the semantic processing of quantifiers disrupted by cognitive load? Experiment 2 demonstrated that standard scalars were processed very differently from numerals. During the numerals trials, the proportion of looks to the target character reached significantly above chance during the Quantifier region, which did not occur for the standard scalars until the Noun Start region for the standard scalars. It seemed that quantifier information from numerals was used rapidly to resolve referential ambiguity regardless of the cognitive load, whereas quantifier information from standard scalars was integrated more slowly, likely due to a disruption in processing by the cognitive load. This implies that processing standard scalars, even just in terms of their semantic content, was a more effortful process than comprehending numerals. It is worthwhile to note that this phenomenon is separate from the questions of whether scalar implicatures are effortful and how numbers receive their upper-bounds,

which were discussed in prior literature (such as De Neys & Schaeken, 2007; Huang & Snedeker, 2009; Marty et al., in submission).

One hypothesis for this data pattern could reflect the ease with which the meanings of numbers can be retrieved from the lexicon. In general, it has been found that words that occur more frequently are retrieved more quickly (e.g., Cleland et al., 2006). However, according to Google Ngrams, which analyses the frequency of word occurrence in a corpus of books written in a given language (in this case, English), out of these scalar quantifiers, "all" appeared the most frequently in 2008 at 1900 per million unigrams (i.e., one-word strings), followed by "some" at 1000 per million unigrams, "two" at 900 per million unigrams, and "three" at 400 per million unigrams. Thus, it seems that standard scalars are more commonplace in the English lexicon than are numerals, and this data pattern cannot be explained by faster recognition of numerals.

Another plausible explanation for the differential processing cost of numerals and standard scalars is that the mapping from the auditory stimuli during the prompt sentence to the visual scene was more difficult for standard scalars than for numerals. Whereas the applicability of the numerals as a description for the visual stimuli could be verified solely by examining the quantity of objects in each quadrant, that of the standard scalars depended on the quantifier of objects in another quadrant (Huang & Snedeker, 2009). That is, in Experiment 1, participants would have had to consider the quantity of objects that the adjacent character possessed to ensure that "some" or "all" would be appropriate descriptors for the scene. The greater suitability of the numerals as quantifiers in this case could have

contributed to the comparable speed of referential ambiguity resolution in Experiment 2.

The data pattern from these two experiments could also be account for by the notion that semantic composition, which occurs as quantifiers are processed and incorporated into context during sentence comprehension, was slower for standard scalars than for numerals. According to prior literature on language processing, comprehension is incremental, such that lexical information from a word is processed and integrated with ongoing sentence-level representations (Urbach & Kutas, 2010). Researchers have found words that are a poor semantic fit or unexpected in context to elicit a larger N400, which signifies difficulties in semantic processing (Kutas & Hillyard, 1984). Whereas Kaan et al. (2007) showed that the infelicitous occurrence of numerals in context resulted in longer reading times and larger N400s, Urbach & Kutas (2010) found that the semantics of standard scalars had very little effect on these metrics. Specifically, in Urbach & Kutas (2010), a smaller N400 was detected for "Many farmers grow worms; few farmers grow crops" than for "Many farmers grow crops; few farmers grow worms", which was surprising because the former was less plausible than the latter. This is especially notable given that in Kaan et al. (2007), a greater N400 was detected for "Five ships appeared on the horizon; six were bombarded by enemy fire" than for "Five ships appeared on the horizon; two were bombarded by enemy fire", which was in accordance with offline plausibility judgments. This further distinguishes the ways by which numerals and standard scalars are processed, and suggests that standard scalars are comprehended on a more delayed time course than are numerals.

In conclusion, the effect of cognitive load on quantifier processing differs between the two types of scalar quantifiers, standard scalars and numerals. For the former, even the slightest load seemed to disrupt and delay the semantic analysis component of quantifier processing, but for the latter, semantic analysis was robust and rapid across all load conditions. We can speculate about the differences in the mechanism of comprehension between standard scalars and numerals. Though we were not able to address the question with which we began the investigation—how and to what extent cognitive load would affect pragmatic inferencing in quantifier processing—this thesis can still contribute to the ongoing discussion on the possible systematic processing differences among quantifier types.

Appendix A: Table of Load Task Stimuli

Zero/Two-Load Letter Sequences	Four-Load Letter Sequences
B H	B H F J
F J	L R M X
L R	H L X F
M X	R H M L
H L	J H F R
X F	B L J X
R H	R M H F
M L	X H L B
J H	B R F J
F R	H X R L
B R	F H X M
J X	M R H B
R M	X J H R
H F	R B F X
X H	L M J R
L B	J F M B

Appendix B: Table of Eye-tracking Task Stimuli

Item Pair Base	Object A	Object B
birthday	cakes	cards
baseball	bats	gloves
football	helmets	jerseys
Christmas	lights	trees
apple	pies	sauce
music	boxes	stands
micro	phones	waves
motor	boats	cycles
hockey	pucks	sticks
fire	crackers	flies
butter	cups	flies
toilet	scrubbers	paper
coffee	creamers	makers
table	cloths	spoons
honey	bees	dews
water	fountains	melons
photo	copiers	albums
bottle	caps	openers
fire	men	places

References

Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition* 118, 84-93.

Barner, D., Chow, K., & Yang, S. (2009). Finding one's meaning: A test of the relation between quantifiers and integers in language development. *Cognitive Psychology*, 58, 195-219.

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.

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

Breheny, R. (2008). A new look at the semantics and pragmatics of numerically quantified noun phrases. *Journal of Semantics*, 25, 93-140.

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

Chierchia, G. (2004). Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. *Structures and beyond*, 3, 39-103.

Cleland, A.A., Gaskell, M.G., Quinland, P.T., & Tamminen, J. (2006). Frequency effects in spoken and visual word recognition: evidence from dual-task methodologies. *Journal of Experimental Psychology*, 32, 104-119.

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.

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

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.

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

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

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

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

Horn, L. (1972). *On the semantic properties of logical operators in English.* Indiana University Linguistics Club.

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

Huang, Y., & Snedeker, J. (2009b). Semantic meaning and pragmatic interpretation in 5-year-olds: Evidence from real-time spoken language comprehension. *Developmental Psychology*, 45, 1723.

Huang, Y., Spelke, E., & Snedeker, J. (in submission). What do numbers exactly mean?

Hurewitz, F., Papafragou, A., Gleitman, L., & Gelman, R. (2006). Asymmetries in the acquisition of numbers and quantifiers. *Language Learning and Development*, 2, 77-96.

Kaan, E., Dallas, A. C., & Barkley, C. M. (2007). Processing bare quantifiers in discourse. *Brain Research*, 1146, 128–145.

Katsos, N. & Bishop, D. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition*, 120, 67-81.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307, 161–163.

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

Marty, P. & Chemla, E. (2011). Scalar implicatures: working memory and a comparison with 'only'. (Ms. LSCP).

Marty, P., Chemla, E., & Spector, B. (in submission). Interpreting numerals and scalar items under memory load.

Noveck, I. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, 78, 165-188.

Panizza, D., Chierchia, G., & Clifton Jr., C. (2009). On the role of entailment patterns and scalar implicatures in the processing of numerals. *Journal of Memory and Language* 61, 503-518.

Panizza, D., Chierchia, G., Huang, Y., & Snedeker, J. (in press). The Relevance of Polarity for the Online Interpretation of Scalar Terms. To appear in *The Proceedings of Semantics and Linguistic Theory* (SALT) 19.

Papafragou, A., & Musolino, J. (2003). Scalar implicatures: Experiments at the semantics-pragmatics interface. *Cognition*, 86, 253-282.

Pouscoulous, N., Noveck, I., Politzer, G., & Bastide, A. (2007). A developmental investigation of processing costs in implicature production. *Language Acquisition*, 14 (4), 347.

Tanenhaus, M. K., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science,* 268, 1632–1634.

Sauerland, U. (2012). The computation of scalar implicatures: pragmatic, lexical or grammatical*? Language and Linguistics Compass* 6. 36-49.

Sperber, D. & Wilson, D. (2004) "Relevance Theory" in G. Ward and L. Horn (eds) *Handbook of Pragmatics*. Oxford: Blackwell, 607-632.

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

Zondervan, A. (2010). Scalar implicatures or focus: an experimental approach (Vol. 249). LOT Dissertation Series.