What exactly do numbers mean?

Yi Ting Huang, Elizabeth Spelke, and Jesse Snedeker
Harvard University

Acknowledgments. We would like to thank Julien Musolino for introducing us to these issues, Susan Carey who kept asking us the right questions, and David Barner who has always kept us on our toes. This work also benefited from conversations with Anna Papafragou, Gennaro Chierchia, Steven Pinker, and Jeff Lidz. We are grateful to Barbara Sarnecka who shared her data with us and Sasha Yakhkind who assisted in data collection. This research was supported by National Science Foundation Grants 0623845 to JS and 0337055 to ES. Authors' address: Department of Psychology, William James Hall, 33 Kirkland Street, Cambridge MA 02138. Email address: huang@wjh.harvard.edu

Abstract

Number words are generally used to refer to sets of an exact cardinal value, but cognitive scientists disagree about their meanings. Although most psychological analyses presuppose that numbers have exact semantics (*two* means EXACTLY TWO), many linguistic accounts propose that numbers have lower-bounded semantics (AT LEAST TWO), and that speakers restrict their reference through a pragmatic inference (*scalar implicature)*. We address this debate through studies of children. First, we examine toddlers' offline comprehension of number words and find they interpret numbers exactly even in contexts where implicatures are cancelled. Next we consider arguments that patterns of numerical acquisition support lower-bounded semantics and find that these acquisition patterns are better explained by an exact account. Finally, we revisit data from online comprehension and demonstrate that five-year-old children access numerical upper bounds as rapidly as lower bounds, contrary to the hypothesis of lower-bounded semantics. We conclude that number words have exact meanings.

1. Introduction

Where does language end and communication begin? The relationship between semantic and pragmatic interpretation has been a perennial puzzle both in psychology and linguistics. General agreement about the existence of these levels of representation contrasts with controversy over their boundaries. The current paper examines these questions by exploring the controversial case of number words. Linguists have long noted that number words appear to have two distinct interpretations (Horn, 1972 & 1989; Gadzar, 1979; Levinson, 1983 & 2000). Although numbers are often interpreted as specifying exact cardinal values, they can be used in some contexts where the total quantity of items is greater (lower-bounded or "at least" interpretations). For example, in sentence (1) *two* is interpreted exactly. In contrast, in example (2), David uses *two* to mean something like AT LEAST TWO AND POSSIBLY MORE. His statement is true, and felicitous, even if there is a total of 5 chairs in his office.

(1) A bicycle has two wheels, while a tricycle has three. (based on Horn, 1989: 251)
(2) Bonnie: I need to borrow two chairs. Do you know where I could get them?
David: Sure, I've got two chairs in my office. (Kadmon, 2001)

The fact that number words can be interpreted in both of these ways creates a challenge for accounts of number word semantics. Many linguists have suggested that utterances like (2) reveal the lower-bounded semantics of numbers and that exact interpretations only arise through pragmatic inferences (Horn, 1972 & 1989; Gadzar, 1979; Levinson, 1983 & 2000). In contrast, most psychologists assume that number words have exact meanings (Gelman & Gallistel, 1978; Dehaene, 1997; LeCorre & Carey, 2007). This analysis is shared by another class of linguistic theories which propose that number words have an exact semantics and lower-bounded interpretations arise through pragmatic processes (Carston, 1998; Breheny, 2008). Both types of theories argue that the mapping from semantics to ultimate interpretation is complex, suggesting that we cannot trust our pre-theoretical intuitions to guide us to the underlying meaning of these terms.

The present paper explores these two theories of number word semantics by examining the development and comprehension of these terms. In the remainder of the Introduction, we flesh out the two accounts of number semantics, examine the reasons why data from children might be particularly revealing, and briefly look at other recent studies on children's interpretation of number words (see Musolino, 2004 for an earlier discussion of these issues). In the second part of this paper, we examine how number words are interpreted in an offline task, both by adults and by young children who are in the process of acquiring the mapping between number words and quantities. Next, we discuss a recent argument for lower-bounded semantics based on patterns in the early acquisition of number words (Barner & Bachrach, in press) and provide an alternative interpretation of these acquisition patterns. Finally, we revisit data from recent studies of real-time comprehension that, we argue, speak decisively to the current debate (Huang & Snedeker, 2009 & in press).

1.1. Two means AT LEAST TWO: the proposal for Lower-bounded semantics

The issue of number word semantics came to prominence when Horn (1972) argued that the interpretation of numbers closely parallels the interpretation of *scalar terms*: sets of words that

can be arranged in an ordinal relationship with respect to the strength of the information they convey. For example, *some* is part of a scale that includes the stronger term *all*, and *warm* is situated within a scale that also includes *hot*. Scalar terms are typically interpreted as having both an upper and lower bound, giving rise to an interval reading which parallels the exact reading for number words. Thus a sentence like (3) will generally be taken to imply that Henry ate some, but not all, of the ice cream.

(3) Henry: I ate some of the ice cream.
(4) Eva: Did anyone try the lutefisk?
Karl: Yeah, Leif ate some of it. In fact, he ate all of it.

However, in certain contexts scalar terms, like number words, also take on lower-bounded interpretations. Thus in (4), Karl asserts that Leif ate both *some* and *all* of the lutefisk (an infamous Norwegian dish made of fish soaked in lye), indicating that *some* in this context has a meaning which does not exclude the stronger term *all*.

Formal treatments of natural language have generally treated phenomena like these as examples of a pragmatic inference called *scalar implicature* (Horn, 1972; Gadzar, 1979). Following Grice (1957/1975), Horn argued that weak scalars like *some* do not have a semantically-encoded upper bound and therefore are compatible with stronger scalar terms like *all*. Scalars can receive an upper-bounded interpretation, as in (3) via a process of pragmatic inference. This inference is motivated by the listener's implicit expectation that the speaker will make his contribution to the conversation "as informative as is required" (the Maxim of Quantity). For example, if Henry had polished off the ice cream, (5) would be a more informative utterance than (3).

(5) Henry: I ate all of the ice cream.

However, since the speaker did not use this stronger statement, the listener can infer that he does not believe it to be true. Although scalar implicatures are robust across many contexts, they are by definition not a part of the truth conditional content of the sentence and can be cancelled, resulting in overt lower-bounded utterances such as (4).

The Lower-bounded account of number semantics capitalizes on these parallels, arguing that numbers are simply another set of scalar terms. Like other scalars, they have a lower-bounded semantics (*two dogs* means AT LEAST TWO DOGS) but receive an upper bound via scalar implicature. Implicatures are calculated in most situations and thus listeners access the exact interpretation of the utterance. But when implicatures are cancelled, the true meaning of the number word is visible, yielding the marked lower-bounded interpretation as in (2).

Theories which posit a lower-bounded semantics for numerical phrases come in two kinds. Some accounts take the form described above: the lower-bounded meaning arises because the lexical item itself lack an upper bound (Horn, 1972 & 1989; Gadzar, 1979; Levinson, 2000; Winter, 2001). In other theories, the lexical item may have an upper bound but the entire phrase (the determiner phrase or quantifier phrase) generates a mandatory lower-bound meaning as part of semantic composition (Fox & Hackl, 2004; van Rooy & Shulz, 2006; Ionin & Matushansky,

2006; Chierchia, Fox, & Spector, 2008; Barner & Bachrach, in press; Foppollo, Guasti, Chierchia, under review; Panizza, Chierchia, & Clifton, in press). For example, Ionin and Matushansky (2006) argue that cardinal numbers are modifiers which characterize the exact cardinality of sets. Thus, informally, *two* means something like "having exactly two members." However, when numbers compose with nouns and appear in an argument position (e.g., as the subject of a sentence or object of a preposition or verb), a semantic operation occurs which gives them a lower-bounded meaning.

Our data and arguments will speak to both versions of this hypothesis. In all of our experiments, we test participants' understanding of number words in count phrases that occur in argument positions. In these locations, both compositional theories and lexical theories posit a lower-bounded semantics. We also directly compare the interpretation of number words to the interpretation of *some,* a scalar term which is said to produce exact interpretations via the same mechanism as number words according to both versions of the lower bounded hypothesis.

For simplicity of exposition, we will refer to all theories that posit lower-bounded semantic analyses for numerical quantifier phrases as theories in which *number words* have lower-bounded meanings. However, in the General Discussion we will return to this issue and explore whether compositional accounts of lower-bounded semantics could be modified to account for our findings.

1.2. Two means TWO: Theories of Exact semantics

While the Lower-bounded account neatly captures the parallels between cardinal numbers and scalar quantifiers, it flies in the face of the pre-theoretical intuition that numbers have exact meanings. Several theorists have pursued this intuition, arguing that numbers, unlike scalar quantifiers, have exact semantics which delimits both their upper and lower boundaries (Saddock 1984; Koenig, 1991; Horn, 1992; Scharten, 1997; Breheny, 2008; Bultinck, 2005). These accounts have gained support from evidence that numbers pattern differently from other scalar terms. For example, even in semantic contexts in which other scalars systematically receive lower-bounded readings as in (6), the exact interpretation is favored for numbers as in (7).

(6) Everyone who ate some of their berries felt fine.
(7) Everyone who ate two of their berries felt fine.

In these examples, the scalar term appears in the restrictor of the quantifier. Calculating the scalar implicature would narrow down the set of people who feel fine, thus resulting is a weaker statement. For this reason, most pragmatic theories predict that implicatures are typically cancelled in these contexts (Levinson, 2000; Chierchia, Spector, Fox, 2008; Noveck, Chevaux, Guelminger, Sylvestre & Chierchia, 2002; Chierchia, Crain, Guasti, Gualmini, & Meroni, 2001; Panizza et al., in press). Consistent with this prediction, the most natural reading of (6) is one in which *some* is lower bounded, allowing us to conclude that even the folks who ate *all* of their berries felt fine. In contrast, Breheny (2008) argues that in sentences like (7), the number word continues to get an exact interpretation. The statement is accurate and pragmatically felicitous so long as all the people who ate exactly 2 berries are feeling fine. Those who ate 3 *might* feel fine or they *might* be ill, depending on contextual factors, like whether the berries are thought to be poisonous or have beneficial properties. This pattern of interpretation is surprising if number

words receive their upper boundaries through scalar implicature but predictable if these upper boundaries are part of their meanings.

Similarly, Horn (1992) pointed out that yes/no questions with scalar terms can be answered positively when the response is consistent with the lower-bounded semantics of the term but violates the implicature (8). However, when a number is used the most natural response is negative (9).

(8) A: Do some of your friends have children?
B: Yes, they all do. / ?No, they all do
(9) A: Do you have three children?
B: No, I have four/ ?Yes, I have four

Again this pattern would be expected if the upper bound is part of the meaning of the count phrase.

The challenge for an Exact semantics account is to explain how exact meanings can give rise to sentences that appear to have lower-bounded interpretations such as (2). The most common answer is that apparent lower-bounded interpretations arise through another pragmatic process (Carston, 1998). For example, Breheny (2008) suggests that numbers refer to the exact numerosity of a set, but that pragmatic factors play a role in determining *which set* is under discussion. This is implemented by an implicit restrictor that specifies the domain over which quantification occurs. Such restrictors are necessary to explain a variety of phenomena. For example, typically we understand sentences like (10) and (11) as quantifying over some contextually-salient group of people. Without this implicit restrictor, they would quantify over all animate entities and thus be blatantly false.

(10) Everybody came to Allison's party.
(11) Nobody rode the subway; instead they got a ride from Susan.

In many contexts, count phrases quantify over the all entities of the relevant kind that are available in a given context, giving rise to exact readings like those in (1) and (9). Because larger sets necessarily contain smaller sets, however, further implicit restriction of the domain will create what appear to be lower-bounded interpretations.

Contextual factors can increase or decrease the probability of further domain restriction. In general, it is most relevant to know the total number of items in the context and thus the exact reading is favored. However, in certain circumstances, it is more pertinent to the goals of the conversation to ascertain whether a subset of a particular numerosity exists. In (2), for example, Bonnie is more interested in whether David has *any* set of 2 chairs than she is in learning the total number of chairs in his possession. Under these circumstances, the number still has an exact meaning but gives rise to an apparent lower-bounded interpretation because the quantified phrase refers to some subset (of that exact numerosity) rather than the maximal set.[1]

[1] This account is complicated by the fact that in cases such as (2) the count phrase does not appear to refer to any *particular* set of 2 chairs. Breheny (2008) argues that the same problem

To summarize, both Lower-bounded and Exact semantics accounts provide prima facie adequate explanations for the dominance of exact interpretations of number words and the occasional appearance of lower-bounded interpretations. Furthermore, both accounts do so by making use of the distinction between semantic representations and pragmatic processes. The Lower-bounded account does so by stating that, like other scalar quantifiers, number words have lower-bounded meanings (*two* means AT LEAST TWO) and receive exact interpretations via the pragmatic inference of scalar implicature. Alternatively, the Exact semantics account states that number words have exact meanings (*two* means EXACTLY TWO), which are evident when they refer to the maximal set (Breheny, 2008), while apparent lower-bounded interpretations are attributed to pragmatic factors that lead to set decomposition.

1.3. Why children's interpretation of numbers might be particularly informative

Several researchers have suggested that we might gain a better understanding of the semantics of number words by looking at how they are interpreted by young children. Children are notoriously poor at calculating scalar implicatures (Paris, 1973; Smith, 1980; Braine & Rumain, 1981; Noveck, 2001; Papafragou & Musolino, 2003; Chierchia et al., 2001; Gualmini, Crain, Meroni, Chierchia, & Guasti, 2001; Barner, Chow, & Yang, 2009; Huang & Snedeker, in press; Foppollo et al., under review). In many cases, their pragmatic failures make the lower-bounded semantics of scalar terms visible in their overt judgments. For example, Noveck (2001) found that seven- to ten-year-olds, unlike adults, often interpreted weaker modal statements (e.g. x *might* be y) as being compatible with stronger statements (e.g. x *must* be y). Thus if numbers are upper-bounded only via scalar implicature, as the Lower-bounded theory contends, we might expect that children would accept lower-bounded interpretations, even in contexts where adults prefer exact interpretations.

This issue was first explored by Papafragou and Musolino (2003) who tested both five-year-old children and adults using a pragmatic judgment task. Consistent with previous research, they found that children, but not adults, were content to accept weak scalar predicates (*started*) and quantifiers (*some*) in situations where the stronger scalar term applied (i.e. *finished* or *all*). In contrast, children treated numbers in an adult-like manner, refusing to accept statements like *"Two of the horses jumped over the fence,"* in a context in which they saw exactly 3 horses jump. They concluded that while children often have difficulty computing implicatures, they readily assign exact interpretations to number words. This conclusion receives further support from Hurewitz, Papafragou, Gleitman, and Gelman (2006) who asked three- and four-year-old children to find pictures in which *"The alligator took two of the cookies."* Children, like adults, selected only the picture in which the character had exactly 2 of the 4 cookies, rejecting the one in which he had all 4. However, these same children, unlike adults, happily selected both pictures when asked a parallel question about *some*.

Both of these studies suggest that number word interpretation does not follow the same developmental trajectory as the interpretation of true scalar quantifiers like *some*. However, these studies do not provide clear evidence for the underlying semantics of number words. As

---

arises in the interpretation of specific indefinites and that some of theories developed in that context (Kratzer, 1998; Schwartzchild, 2002) can be extended to the numerical examples.

we saw earlier, both the Exact and the Lower-bounded theories can account for the existence of exact and lower-bounded interpretations, and both theories predict that the exact interpretation will frequently be favored. Evidence that children typically entertain exact interpretations of numbers does not, by itself, tell us how they reach these interpretations. More specifically, this evidence is compatible with two quite different accounts:

Account #1: Exact semantics. On this proposal, children's lower-bounded interpretation of true scalar terms like *some* is taken as evidence of a profound and global difficulty in calculating scalar implicatures. If children cannot calculate scalar implicatures, then their exact interpretations of number words cannot be attributed to this pragmatic process. Thus this strong preference for exact interpretations of numbers must reflect the semantic properties of these terms. This interpretation is called into question, however, by the evidence that young children succeed at calculating scalar implicatures when they are given instructions and training emphasizing pragmatic interpretation over literal truth (Papafragou & Musolino, 2003) and when experimental tasks more closely approximate the role of implicatures in communicative interactions (Papafragou & Tantalou, 2004; Pouscoulous, Noveck, Politzer, & Bastide, 2007; Katsos & Bishop, 2008).

Account #2: Lower-bounded semantics. If implicature is variable in childhood rather than absent, then the discrepancy between true scalars and numbers could reflect differences in the pragmatic processing of these items rather than differences in their meanings. On this hypothesis, children learn to calculate upward-bounding implicatures for *two* earlier than they learn to calculate them for *some* or *start*. This precocity could be fueled by several factors: the frequency with which particular implicatures are suspended, greater contextual support for the upper-bounded interpretation of numbers, parental feedback about the correct use of number words, and the role of the counting routine in allowing children to generate and compare expressions using alternative numerical terms (Papafragou & Musolino, 2003; Barner & Bachrach, in press; Foppolo et al., under review). For example, Barner and Bachrach (in press) have argued that children fail to infer that *some* implies NOT ALL because they fail to spontaneously retrieve and consider alternative expressions including *all* when they encounter *some* (see also Gualmini et al., 2001; Reinhart, 1999; Pouscoulous et al., 2007). In the case of the number words, however, children acquire these terms in the context of a list that is rehearsed extensively and explicitly highlights the alternatives on the scale. Thus, a child who hears the expression *two wheels* may retrieve the stronger alternative, *three wheels,* and infer that the speaker's failure to use this alternative implies that his use of *two* excludes THREE and higher numbers.

This argument suggests that to understand children's interpretation of number words, we must also understand the pathway by which children acquire these words. Young children first produce number words in the context of the counting routine and by two to three years of age, they typically can accurately recite the count list up to *ten* (Gelman & Gallistel, 1978; Wynn, 1990). However, previous developmental research has shown that these words are only mapped onto quantities slowly and sequentially during the preschool years (Sarnecka, Kamenskaya, Yamana, Ogura, & Yudovina, 2007; Le Corre & Carey, 2007; Condry & Spelke, 2008). Wynn (1992) documented this progression by asking children to give the experimenter different numbers of objects from a larger set in the "Give-N" task. In her study, the youngest children

began with a consistent and reliable interpretation of *one*; when asked for *"one fish,"* they would hand the experimenter exactly 1. However, the children failed to show any consistent interpretation for larger numbers. When asked for *two* or more they would always give more than 1 item, but the quantity produced was variable and apparently unrelated to the number requested (see Sarnecka & Lee, 2009). By 2.5 years, most of the children were "two-knowers": they gave the experimenter exactly 2 fish when asked for *two*, but continued to grab a handful of 3 or more when asked for larger quantities. Several months later, these children began responding consistently to *three* ("three-knowers") and by around four-years of age, many had mastered *four* along with the ability to apply their counting routine to enumerate even larger sets.[2]

This prolonged period of limited competence could have profound effects on children's interpretation of known numbers, because scalar implicature critically depends on knowledge of what the speaker might have said. According to the Lower-bounded account, we interpret *two* as EXACTLY TWO because we know that a cooperative speaker could have said *three* if the situation had warranted it. But it's not clear that a two-knower has learned enough about the meaning of *three* to support such an inference. In the absence of a stronger term to drive the implicature, the Lower-bounded theory predicts that the underlying lower-bounded semantics of the term would guide its use. Levinson (2000: 90) developed just such an argument in reference to languages that only possess a small and finite set of number words (Pica, Lemer, Izard, & Dehaene, 2004; Gordon, 2004).

> The scalar prediction is clear in these cases: we have a finite scale <'three', 'two', 'one'>, where 'one' or 'two' will implicate *ceteris paribus* an upper bound; but because there is no stronger item 'four', the cardinal 'three' should lack this clear upper bounding by GCI <Generalized Conventional Implicature>.

Consistent with the logic of scalar implicature, Levinson noted that speakers of these languages often use their largest number word to denote even greater quantities.

For Levinson (2000), the interpretation of number words in young children is also of interest because they have had little or no contact with formal mathematics (see also Geurts, 2006). He suggests that while number words in natural languages have a lower-bounded semantics, educated adults (and presumably school-aged children) have also acquired exact meanings for the mathematical numerals through formal education. On this theory, numerals and number words are homophones for mathematically literate adults, and it is difficult for such informants to separate them when making judgments. Because two- and three-year-old children have had little direct exposure to formal mathematics, however, they are unlikely to be influenced by this putative lexical ambiguity. These two considerations suggest that Exact and Lower-bounded theories of number word semantics may best be distinguished by examining their interpretations in very young children who have mastered the meanings of some, but not all, of the numbers in their count list.

1.4. How can we discover what number words mean?

---

[2] The alert reader might wonder whether performance on the Give-N task might reveal more about the nature of early number semantics. An analysis of this kind forms the core of Barner and Bachrach's (in press) defense of lower-bounded semantics and is addressed in the third section of this paper.

In order to truly isolate the meaning of number words, we must adopt paradigms that disentangle the semantics of the terms from the contributions of pragmatic implicatures. Because previous judgment tasks have used contexts in which adults typically calculate implicatures, it is not clear whether children's exact number-word interpretations in these contexts reflect an exact semantics or a lower-bounded semantics supplemented by upward bounding implicature (Papafragou & Musolino, 2003; Hurewitz et al., 2006). To determine the semantic meanings of number words, we need to test participants in contexts in which implicatures are cancelled. Our judgment task was designed to do this.

Lower-bounded theories predict that exact readings will arise whenever the scalar implicature is calculated. On some versions of this hypothesis, implicatures are calculated by default (Levinson, 2000) and thus the lower-bounded interpretation should only be preferred in tasks in which the implicature is cancelled or by populations who are unwilling or unable to make this inference. We sought to eliminate scalar implicatures in three ways. First, we tested two- and three-year-old children, an age at which scalar implicature appears to be weak or absent. Second, we tested children on the largest number word for which they had a stable mapping, using Wynn's Give-N task to identify children who displayed mastery of *two* but not *three* and then assessing their interpretation of *two*. We reasoned that, on the Lower-bounded theory, children who lack knowledge of *three* should fail to predict that a cooperative speaker would use *three* rather than *two* to designate sets with more than 2 members. Finally, we designed tasks in which *even adults* would be forced to suspend scalar implicatures.

Crafting a task in which adults suspend scalar implicatures is a challenge, because of the demand characteristics of typical comprehension tasks. When adults are given a verbal description and asked to choose between two or more potential referents, there is a strong task demand to select one and only one of the options provided. For the present question this creates a quandary: Should one of the choices match the exact (or upper-bounded) meaning of the term? If an exact match is not included (for *two* the participant is given a choice between 1 and 3), then any person who interprets the term exactly has no valid response and thus must either protest or reconstrue the task in some way. However, if an exact match is included (e.g., for *two* a choice between 2 and 3) then the very presence of this contrast could promote an implicature even if both possibilities are consistent with the semantics of the term. In fact, Hurewitz and colleagues (2006) found that adults systematically make implicatures for *some* when both a lower-bounded referent and upper-bounded referent are present.

To reduce this demand, we created a context in which scalar implicatures would be cancelled while simultaneously providing participants with a clear way to indicate an exact interpretation of a critical word: a decoy was provided that could be interpreted as an exact match, or not, as the participant saw fit. Thus participants were asked to *"Give me the box with two fish"* in the context of a visible mismatch (e.g., a box with 1 fish), a visible and salient lower-bounded target (e.g., a box with 3 fish) and a covered box with unknown contents. By using boxes, we delineated each set with a clear physical boundary and forced the participant to act upon it as a whole rather than as a collection of individual units. On the Exact semantics theory, this should make it difficult to decompose the set into smaller subsets, minimizing the possibility of a lower-bounded interpretation arising from a pragmatic shift in the domain of reference (e.g., from the 5 chairs in David's office to the 2 chairs that Bonnie needs; Breheny, 2008). Furthermore, the

critical instruction embeds the count phrase in a single definite description (*"the box with two fish"*). This leads to a presupposition that there is one unique referent in the context that satisfies the description (Ferge, 1892; Strawson, 1950), as well as increasing the likelihood that the count phrase would be mapped to the entire set.

By definition implicatures can be cancelled and we reasoned that any upper-bounding implicature would be cancelled when it is in conflict with other information provided by the speaker. Specifically in this context, if the word in question (*two*) has lower-bounded semantics, then the presupposition of the definite description would be unambiguously satisfied by the visible lower-bounded match (box with 3 fish). Since this referent was accessible to both parties and the contents of the covered box were not, we expected that participants would cancel the scalar implicature, bending the maxim of quantity to avoid an egregious violation of the maxim of quality (Grice, 1975). However, if the word in question has an exact semantics, this option would not be available: without a visible match, the participant would have to conclude that the referent is in the covered box.

To ensure that this task was successful in canceling the implicature, we tested our methods on the perennial poster child for scalar implicature, the quantifier *some*. On the critical trials, participants were presented with a box where Cookie Monster had none of the cookies, another box where he had all of the cookies and a covered box, and we asked them to "*Give me the box where Cookie Monster has some of the cookies.*" If the covered box task succeeds in canceling the implicature, then both children *and* adults should select the box where he has all the cookies.

2. Interpretation of scalar terms and number words in early acquisition

2.1. Scalar conditions: When are scalar implicatures cancelled?

2.1.1. Methods

Thirty English-speaking undergraduates from Harvard University, and 10 English-speaking children between the ages of 2;6 and 3;5 (mean 2;9) participated in the experiment. The different numbers of participants in the two age groups reflected a difference in experimental design. In children, the trial types were manipulated within subjects to provide more information about each child's understanding of the task and the scalar terms. In contrast, in adults, the trial types were manipulated between subjects to ensure that participants did not use information from one trial to draw inferences about the contents of the covered box in other trials. Thus ten adults and ten children were tested on each trial type.

The study was composed of three parts. During the Pretest phase (conducted only for children), we used a modified version of the Give-N task to elicit knowledge of *some* and *all*. Children were presented with several small plastic fish and were simply asked to "*Put some (all) of the fish*" into a basket ("the pond"). All children who were tested demonstrated knowledge of these scalar terms by putting at least 1 fish in the basket when asked for *some* and by putting the entire quantity when asked for *all*.

During the Familiarization phase, we introduced participants to the covered box task. On each trial they were presented with two open boxes containing toy animals and a third covered box and were asked to give the experimenter the box that contained a particular animal. The

target animal was in one of the open boxes on two of the familiarization trials and hidden inside of the covered box on the remaining two. This sequence of four trials was repeated twice. The first time participants were given feedback after each choice and were allowed to open the covered box in their search for the target animal. The second time through, they were told not to open the covered box and they were not given any feedback. All participants selected correct boxes when the target animal was visible and selected the covered box when it was not and were therefore included in this experiment.

During the Test phase, participants were presented with boxes that contained pictures. In each picture there were two characters (Cookie Monster and Big Bird) and a set of cookies that belonged to one of them or was split between them. Participants were asked to "*Give me the box where Cookie Monster has some of the cookies*" in the following three contexts (see Figure 1):

INSERT FIGURE 1 ABOUT HERE

1. NONE vs. SOME. On these trials, participants were presented with three boxes: one empty set match (a box with a picture where Cookie Monster had none of the cookies and Big Bird had all of them), one subset match (where both Cookie Monster and Big Bird have some but not all of the cookies), and a third covered box.

2. ALL vs. SOME. On these trials, participants were presented with one total set match (a picture where Cookie Monster had all of the cookies and Big Bird had none of them), one subset match (another picture where both had some but not all of the cookies), and a third covered box.

3. NONE vs. ALL. On these trials, participants were presented with one empty set match (a picture where Cookie Monster had none of the cookies and Big Bird had all of them), one total set match (a picture where Cookie Monster had all of the cookies and Big Bird had none of them), and a third covered box.

The three boxes were presented side-by-side in random linear order. There were three tokens of each trial type, each featuring different pairs of characters sharing various objects (e.g., Winnie-the-Pooh and Piglet with apples, Barney and Tinky-Winky with lollipops). Trials were presented to children in a pseudo-randomized order which varied the presentation order of the three character pairs and the order of the three trial types within these blocks. Each adult received just three trials, all of the same type. This ensured that their responses reflected their naïve understanding of the sentences rather than any inferences they might draw about the goals of the study by comparing the different trials types.

We had three distinct predictions. First, in the NONE vs. SOME trials, children and adults should consistently pick the subset match since it is the only visible option that is consistent the semantics of the quantifier. Second, in the ALL vs. SOME trials, adults should calculate the implicature and select the subset match, but children should fail to do so, choosing either the subset or total set match at random as in past studies (e.g., Papafragou & Musolino, 2003; Hurewitz et al., 2006; Foppollo et al., under review). Finally, the critical NONE vs. ALL trials test our hypothesis that scalar implicatures will be cancelled when a definite description is used in

the absence of a clear implicature match but in the presence of a set which matches the lower-bounded semantics of the term. If implicatures are cancelled, then both children *and adults* will select the total set match. Selecting the covered box would suggest that participants had calculated the implicature, rejected total set match, and inferred that a subset match must be present in the mystery box.

2.1.2. Results.

Figure 2 indicates that the response pattern varied across the three types of trials and across the two age groups. Both children and adults in the NONE vs. SOME trials overwhelmingly selected the box where Cookie Monster had a subset of the cookies (M = 93% and M = 100%), with no difference between the two age groups (W = 95, Z = .76, p > .4). Thus the children clearly understood the task and recognized that *some* is incompatible with *none*. In ALL vs. SOME trials, however, adults overwhelming favored the box with the subset of the cookies (M = 90%) whereas children were equally disposed to select the box containing the subset of cookies and the box containing the total set (M = 50% and M = 40% respectively). The proportion of subset selections differed reliably between the two groups (W= 69.5, Z = 2.68, p < .01). In fact, comparisons to chance revealed that children did not have a reliable preference for the subset match (t(9)=.60, p > .5), unlike adults (t(9)=5.61, p<.001). These results are consistent with previous studies demonstrating that adults calculate scalar implicatures when an implicature match is present whereas young children often do not.

INSERT FIGURE 2 ABOUT HERE

Critically, on the NONE vs. ALL trials, both the children and the adults strongly favored the box containing the total set of cookies (M = 83% and M = 87% respectively, W = 96.5, Z = .64, p > .5). Very few selected the covered box (M = 7% and M = 13% respectively, W = 100, Z = .73 p > .5). In the absence of a visible implicature match, therefore, even adults accepted the box where Cookie Monster had all of the cookies, rather than questioning the request or inferring that the covered box must contain the subset option. This finding strongly suggests that our contextual manipulation succeeded in canceling the scalar implicature.

2.1.3. Discussion.

The results of our covered box task both confirm and extend prior studies examining the interpretation of *some* by adults and children. Like other researchers, we found that children often fail to calculate scalar implicatures in contexts where adults make these inferences (Hurewitz et al., 2006; Papafragou & Musolino, 2003; Foppollo et al., under review). Specifically, when asked for *"some of the cookies"* in the ALL vs. SOME trials, children split their choices between the boxes containing the subset and the total set, demonstrating a failure to use scalar implicature to restrict the reference of the definite description. Adults, on the other hand, overwhelmingly favored the subset match. However, we extended this work by demonstrating that when no clear match for the scalar implicature is provided, both adults and children happily select the lower-bounded quantity (i.e. all of the cookies) demonstrating a common semantic representation of the scalar term. This finding supports our conjecture that scalar implicatures are cancelled when the critical terms are used as part of a definite description in the presence of a salient and accessible lower-bounded alternative and the absence of a clear implicature match.

This interpretation of our data is consistent with the prior literature on the interpretation of *some* and the development of scalar implicature. Nevertheless, we considered the possibility that the children's response pattern reflected simpler strategies that were unique to this experiment. First, children may have simply ignored the quantifier and picked a card in which the target character was associated with cookies. Second, children may have failed to understand that each box was to be evaluated in isolation and instead interpreted the quantifier with respect to the contents of all three boxes. On this interpretation there are two different Cookie Monsters and two different Big Birds splitting up cookies and since there are cookies in both visible boxes, no character has all the cookies.

To rule out these possibilities we conducted an additional study in which children were asked to select the box where Cookie Monster had *"all of the cookies."* The experiment was identical in all other respects to the scalar task and ten children between 2;6 and 3;5 participated. If children ignored the quantifier we would expect the same pattern of performance with *all* that we saw with *some*. If children quantified across the three boxes, they should either refuse to answer the question (no one has all the cookies) or perform at chance. But they did neither of these things. On the critical ALL vs. SOME trials, children who were asked for *all* systematically selected the card in which Cookie Monster had all the cookies (93%), while children who were asked for *some* split their responses between the some and all cards ($p < .01$). On the ALL vs. NONE trials, the children consistently selected the card where Cookie Monster had all the cookies and Big Bird had none, despite the fact that the Big Bird on the other card also had some cookies ($M = 93\%$). Thus the children's pattern of performance for *all* demonstrates that they are able to use semantic information about the quantifier to reliably select the correct box.

Thus performance in the scalar conditions demonstrates that implicatures are cancelled in the critical trials of the box task; when the exact match is absent, even pragmatically sophisticated adults accept the lower-bounded alternative. In the next section, we used this task to probe interpretations of number words in the same two populations: adults and children who displayed mastery of *two* but not *three*. During the critical test trials, participants were shown similar box displays featuring different numbers of objects: a box with 1 fish, another box with 3 fish, and a covered box, and they were asked to *"Give me the box with two fish."* If numbers have an exact semantics, then participants should select the covered box, inferring from the definite description that a set of exactly 2 fish must be in there. If, on the other hand, numbers have a lower-bounded semantics, then participants should select the box with 3 fish, consistent with their willingness to accept the total set as an instance of *some* in the parallel condition in the scalar quantifier experiment.

2.2. Number conditions: What are the underlying semantics when implicatures are cancelled?

2.2.1. Methods

Thirty English-speaking undergraduates from Harvard University and 10 English-speaking children between the ages of 2;6 and 3;5 (mean 3;0) participated in this experiment. We selected only those children who knew the meaning of *two* but did not know the meaning of *three* (two-knowers). Four children were excluded because they did not meet these criteria. As in the previous experiment, children were tested on all three types of trials, whereas separate groups of

adults were tested on each type of trial to prevent across-trial inferences about the contents of the covered box.

Wynn's (1992) Give-N task was used as a pretest to determine the level of number word knowledge by asking children to put different quantities of fish into a basket. Children were classified as two-knowers if they gave 1 fish when asked *one*, 2 in response to *two*, and an arbitrary larger number in response to all other requests. The first ten children that we identified in this group participated in this study. We also elicited children's knowledge of the count list and found that all participants were able to count up to *ten*.

The materials, design and procedure of the number condition were similar to those of the scalar conditions. In the Familiarization phase, participants were introduced to the box task using the procedure from the previous task. Only participants who selected correct boxes when the target animal was visible and selected the covered box when it was not were included in this experiment. One child was excluded for failing to meet this criterion. During the Test phase, participants were asked to *"Give me the box with two fish"* in the following contexts (see Figure 3):

INSERT FIGURE 3 ABOUT HERE

1. EXACT VS. LESS. During these trials, participants were given a box containing 2 fish (exact match), a box containing 1 fish (the less-than option), and a covered box. This condition corresponded to the "SOME vs. NONE" trials in the scalar condition.

2. EXACT VS. MORE. During these trials, participants were given box containing 2 fish (exact match), a box containing either 3 or 5 fish (the more-than option), and a covered box. This condition corresponded to the "SOME vs. ALL" trials in the scalar condition and closely parallels the task used by Hurewitz et al. (2006).

3. LESS VS. MORE. During these critical trials the exact match was absent. Participants saw a box containing 1 fish (the less-than option), a box containing either 3 or 5 fish (the more-than option), and a third covered box. This condition corresponded to the "ALL vs. NONE" trials in the scalar condition, in which children and adults suspended the scalar implicature and chose the visible match to the lower-bound quantifier meaning.

In the EXACT VS. LESS trials, both the Lower-bounded and the Exact theories predict that children and adults will consistently pick the exact match, since it falls within the meaning of the term, while the other visible option does not. In the EXACT VS. MORE trials, the Exact and Lower-bound theories both predict that adults should consistently select the exact match, but they may make diverging predictions for children. The Exact theory predicts that children also will select the exact match, since it is the only choice matching the meaning of the term. The predictions of the Lower-bounded theory depend on our assumptions about the development of scalar implicature and children's knowledge of number words. If implicature develops as a unitary ability, then the Lower-bounded theory predicts that children will select both options randomly, as they did when tested with *some*. If implicature develops in a piecemeal fashion, and is easier for numbers than other quantifiers then the predictions of the Lower-bounded account depend on

the child's knowledge of number words. The children in this study were selected because they did not demonstrate knowledge of the meaning of *three* in the Give-N task. On the Lower-bounded account, knowledge of the larger term is necessary for making the implicature, thus ignorance of the meaning of *three* should prevent them from systematically selecting the exact match (Levinson, 2000), unless the Give-N task systematically underestimates children's knowledge of numbers (see Barner & Bachrach, in press), an issue we return to in section 3.

In contrast, performance on the LESS VS. MORE trials should distinguish between the two accounts, regardless of the participants' numerical knowledge. In the scalar task, we found that when a subset match was absent, both children and adults failed to generate a scalar implicature for *some*. Thus in this context, the Exact and Lower-bounded theories make distinct predictions for performance at *both* ages. The Exact theory predicts that the participants will reject the visible choices, since they are not compatible with the meaning of the expression, and will infer that the correct set is in the covered box. In contrast, the Lower-bounded theory predicts that both adults and children will select the lower-bounded option since it is compatible with the semantics of the number word.

2.2.2. Results

Figure 4 illustrates that participants overwhelmingly preferred to select an exact match when it was visibly present. When asked for *two*, both children and adults chose the box containing exactly 2 fish in both the EXACT VS. LESS trials (M = 95% and M = 100%) and the EXACT VS. MORE trials (M = 93% and M = 100%). We found no differences across the age groups (both p's > .2). Performance in the EXACT VS. MORE trials is notable because it demonstrates a divergence between scalars and numbers in children: in the presence of larger quantity children fail to generate an implicature in the scalar task but show an adult-like preference for the exact interpretation of *two*. This is consistent with previous studies comparing scalar terms and number words in children (Papafragou & Musolino, 2003; Hurewitz et al., 2006; Foppollo et al., under review). However, as we noted above, it is unclear from these data whether children's divergent preferences are a result of an exact semantics for numbers or a greater precocity with scalar implicatures along the number scale.

INSERT FIGURE 4 ABOUT HERE

The critical LESS VS. MORE trials explore this issue directly. On these trials, where there was no exact match, participants consistently selected the covered box (M = 95% and M = 100% for children and adults respectively). Thus there was a reliable effect of trial type; both children and the adults were far more likely to select the covered box on the LESS VS. MORE trials (W = 55, Z = 2.78, p < .005; W = 55, Z = 2.78, p < .001, respectively). There were no reliable differences between children and adults for any of the trial types (all Z's < .8; all p's > .2). Thus even at this early stage of acquisition, the children's interpretation of number words mirrors that of adults.

A comparison of Figures 2 and 4 suggests that participants' responses for *two* were quite different than their responses for *some*. This contrast was confirmed by a series of 2 x 2 ANOVAs comparing differences in performance across the two ages (adults vs. children) and conditions (scalar vs. number). First, we examined the proportion of SOME/EXACT matches in the trials that featured a visible contrast between the SOME/EXACT match (subset/2) and the

ALL/MORE-THAN match (total set/3 or 5). While we found main effects of age ($F(1, 36) = 5.16$, $p < .05$) and condition ($F(1, 36) = 6.63$, $p < .05$), these effects were driven by the presence of an interaction between the two variables ($F(1, 36) = 6.63$, $p < .05$). While adults chose the subset/exact match in both conditions, children failed to restrict their interpretation to the subset for scalar quantifiers but did so for number words.

Next we examined the proportion of covered box choices in the critical trials that featured a NONE/LESS-THAN match (empty set/1) and an ALL/MORE-THAN match (total set/3 or 5) but no visible exact match. While we found no main effect of age ($p > .30$) and no significant interaction between age and condition ($p > .90$), we did find that subjects selected the covered box significantly more in the number task than in the scalar task, $F(1, 36) = 192.24$, $p < .001$. This effect was robust in both children ($t(18) = 139.23$, $p < .001$) and adults ($t(18) = 72.59$, $p < .001$).

2.2.3. Discussion

The results from the number task provide a stark contrast to children's performance with the scalar quantifier *some*. Like adults, children overwhelmingly selected the exact match when it was visible and selected the covered box when it was not. Since the task was to find one box that uniquely satisfies the description (*"the box with two fish"*), a theory of lower-bounded semantics most naturally predicts that participants should select the visible box with more objects without ever needing to consider the covered box. Instead participants rejected the visible lower-bounded option and inferred that the covered box must have 2 fish in it. In contrast, when there was a visible exact quantity match, participants had no difficulty ignoring the covered box and selecting this item.

On the Lower-bounded theory, the rejection of the larger or lower-bounded target can only be explained as the effect of an upward-bounding scalar implicature. There are two features of this study which make such an explanation unlikely. First, the children that we tested in the number task gave exact interpretations of *two* even though they seemed to have little knowledge of the meaning of the word *three*. On the Lower-bounded account, scalar implicature is motivated by mutual knowledge of the terms on the scale and their relative informational strength. Second, on the critical trials, participants were tested in a context in which scalar implicatures appear to be cancelled. In the same task with the scalar quantifier *some*, both children and adults accepted the lower-bounded match (ALL) when no implicature match was visible. When we presented number words in a parallel context, in contrast, our participants consistently rejected the lower-bounded match (3), suggesting that the upper bound was a semantic requirement rather than a pragmatic preference.

The experiment also illustrates the virtues of testing young children and adults simultaneously. Taken by themselves, the data from adults could be interpreted to reflect two distinct processes: a process for treating the quantifiers *all* and *some* as natural language scalar terms, and a distinct process for treating the numbers *two* and *three* as terms in formal mathematics. Although all the terms were presented in a natural language context, it is possible that adults' extensive history of mathematical instruction and experience, in which numbers are used only in their exact meanings, could have biased their interpretations of number words (Levinson, 2000). The children in the present study, in contrast, have received no instruction in

mathematics and have not even mastered the first step of symbolic mathematics, the construction of the count list. The common profile of performance of children and adults therefore casts doubt on Levinson's interpretation of data from adults and provides evidence for a developmentally invariant, exact semantics of number words.

While the present findings appear to provide strong support for exact semantics, this conclusion is challenged by a recent sweeping proposal made by Barner and Bachrach (in press, henceforth B&B). In designing the box experiments, we worked under the assumption that the Give-N task provided an accurate assessment of the largest number word that a child has managed to acquire. B&B have challenged this assumption and provide a cogent argument in favor of a lower-bounded semantics. In the next section, we outline their argument, describe the evidence supporting their account, propose an alternative explanation of this evidence, and present new analyses that begin to distinguish between these two accounts.

3. Weighing the evidence: New arguments for lower-bounded semantics

3.1. Barner & Bachrach's defense of lower-bounded number word semantics

B&B argue that numerically quantified noun phrases have a lower-bounded semantics and that children will only consistently derive an upper-bounded interpretation for phrases with a particular number word once they have learned the next number on the count list.[3] Thus they would predict that a *true* two-knower would accept 3 as an instance of *two* in selection tasks like ours. But critically, they propose that performance on the Give-N task and on most other measures of number word mastery consistently underestimates children's knowledge by exactly one number. For example, they would argue that children who appear to be two-knowers are actually three-knowers and use *three* to provide an upper bound for *two*. Since these children do not yet know *four,* they cannot perform the implicature that would allow them to infer that *three* does not usually refer to sets of 4. Consequently, they allow *three* to refer to larger quantities and thus get classified as two-knowers in Wynn's system. As evidence for this proposal, they note that children's performance on the next highest number in the count list is actually quite systematic. For example, across three experiments using the Give-N task, B&B found that non-, one-, and two-knowers (hereafter, "N-knowers") gave the correct quantity 51% of the time when asked for N+1. This preference also emerged in a task where children gave numerical descriptions for pictures (the "What's-on-this-Card" task). Across two experiments, children used the correct number word for N+1 objects 62% of the time. Thus they argue that children

[3] Specifically B&B adopt the proposal that number words lexically encode exact cardinalities, but numerical quantified noun phrases have a lower-bounded semantics due to an operation of existential closure. As we noted above, for our present purposes this kind of proposal is identical to one in which number words themselves have lower-bounded semantics, because all the evidence under discussion probes the interpretation of numbers in argument positions in which existential closure would operate. Critically, B&B explicitly tie their proposal to the empirical phenomena under investigation here, by arguing 1) that the lowerbounded interpretation of a number should be visible when the next largest number is not know and 2) that the process by which upper bounds are calculated is parallel for numbers and scalar quantifiers. We return to these issues in the General Discussion.

who appear to be N-knowers, have actually assigned the mature, lower-bounded meaning to N+1.

B&B also offer an account of the divergence between children's interpretation of number words and their interpretation of quantifiers like *some*, which has been documented both in our research and preceding experiments (Papafragou & Musolino, 2003; Hurewitz et al., 2006; Foppollo et al., under review). They argue that scalar implicature is not intrinsically difficult for children, but their performance is impaired by difficulty retrieving the scale during comprehension. Scale retrieval is less challenging with number words because the number scale consists of a stable set of alternatives that are laid out in the count list, which children master long before they develop any stable word-to-quantity mappings. B&B argue that encountering one term on this scale leads even the youngest children to retrieve the number list and calculate the relevant upper bound. In fact, they argue that children compute implicatures to interpret numbers from the time they become one-knowers at about 2;6. In contrast, non-numerical scales are rarely if ever practiced, and so children often fail to retrieve them (e.g., not generating *all* when presented with *some*). The salience of these scales increases gradually across development, leading to the steady changes in children's interpretation of a variety of scalar terms.

This hypothesis provides a serious challenge to the conclusions drawn in the previous section. If it is correct, then the children we tested in the box task were not two-knowers, as argued, but three-knowers. By hypothesis they were able to interpret *two* exactly, not because it has an exact meaning but because they already understood that *three* means AT LEAST THREE, and were able to use this term to infer that *two* could not refer to a set of 3 items.

3.2. A counter-hypothesis: Exact semantics and the emergence of knowledge

Do children who interpret *two* exactly have robust enough knowledge of the meaning of *three* to restrict the lower-bounded meaning of *two* by implicature? Here we contrast B&B's new interpretation of the Give-N data with a different proposal: Two-knowers perform above chance on *three* because children have partial knowledge of its meaning. However, to the extent that children know its meaning, this meaning is exact. To the extent that they do not know its meaning, they treat *three* like any other number word larger than *two*: as contrastive with smaller number words but indistinguishable from larger ones.

3.2.1 What is partial knowledge?

To understand how N-knowers could have partial knowledge of N+1, it is necessary to look closely at the procedure used in the Give-N tasks. Typically in this task, the experimenter begins by asking for *one fish* and continues onto higher numbers in a pseudo-random order (Wynn, 1992; Le Corre & Carey, 2007; Condry & Spelke, 2008). When a child fails to produce a correct quantity (e.g., 3 fish when asked for *three*), the experimenter will typically go back to the next smallest number (i.e., *two*), before retesting the child on the number she got wrong. If the child produces the correct quantity in one instance but an incorrect quantity in the other, then she will be tested a third time which serves as a tie-breaker for deciding on the child's knower level.

Given this procedure, there are two ways in which a group of children who are classified as N-knowers could have partial knowledge of the N+1. First, the knowledge could be binary, but the Give-N task could be an imperfect tool for determining whether it is there. Thus it is possible that every child either knows or does not know the meaning of *three*, but occasionally a child who knows *three* messes up once too often and gets categorized as a two-knower in this procedure. Thus the group as a whole might have partial knowledge, even though no individual has partial knowledge. Classification errors of this kind are expected on any theory. In the presence of random performance errors, any task for sorting children into groups will be imperfect. The small number of trials in the typical Give-N study could exacerbate this problem.

Alternately, children may go through a period in development in which their knowledge of the next number-to-quantity mapping becomes gradually stronger. When this knowledge is weak it influences their responses but does not fully determine them, leading to performance that is above chance but below ceiling. The mapping between a number word and its meaning seems so simple that it is hard to imagine that it could only be partially present, but equally simple mappings appear to emerge gradually and become stronger over time (see Siegler, 2007 for a review). For example, children's ability to retrieve arithmetic facts is initially quite variable: the same child, confronted with the same addition problem on the same day, may retrieve the answer on one trial and then count it out on the next (Siegler & Shrager, 1984). Within the verbal domain, Marcus and colleagues have argued that gradual improvement in memory for irregular past-tense forms underlies children's slow retreat from over-regularization errors (Marcus et al., 1992). Similarly, Gershkoff-Stowe's (2002) analysis of naming errors suggests that the connection between a phonological form and a concept gradually becomes stronger as toddlers gain more experience with a particular word. In the case of number words, we can envision two hypotheses about partial knowledge within an individual. On one hypothesis, the weak emergent knowledge is retrieved on only a fraction of the trials, but on any given trial it is either present or absent. On the other hypothesis, partial knowledge is available on every trial and has a probabilistic effect on interpretation of the semi-known number word. In either case, the children's knowledge leads them to select the exact quantity requested, whereas their ignorance leads them to treat the number as unknown. Critically, on both accounts, at no point in the acquisition of number word knowledge, does the child consider that number words have lower-bounded semantics.

The existing data sets are not sufficiently fine-grained to allow us to tease apart these alternative construals of partial knowledge (variability between children, partial knowledge between trials, or partial knowledge within trials). For our present purposes, however, they are equivalent. All three explain why children's performance on N+1 is above chance but considerably worse than their performance on N, without invoking a lower-bounded semantics. In the next sections, we consider the evidence that could distinguish between a partial knowledge view and the lower-bounded semantics proposal of B&B.

3.2.2 How much partial knowledge of N+1 do children have?

B&B's secondary data analyses provide a starting point for assessing the strength of partial knowledge. In the studies they review, the proportion of trials on which children give N+1 when asked for N+1 ranged from about 23% to 75%. As B&B note, this number alone is not

interpretable, because it could reflect a response bias to grab sets of a particular size. From the standpoint of partial knowledge, the critical comparison is between the proportion of times the child gives N+1 when asked for N+1 (the hit rate), and the proportion of times she gives N+1 when asked for N+2 or more (the false alarm rate). We can use these numbers, in combination with a hypothesis about the nature of numerical knowledge, to estimate the degree to which N-knowers understand N+1.

For example, if we assume that on any given trial a participant either retrieves or does not retrieve the meaning of N+1, then we can express both the hit rate and false alarm rate as a combination of what happens when the child retrieves the word and what happens when she fails (see (12) and (13)).

(12) Hit Rate = (Probability Knowing * 1) + (Probability Not Knowing * Default Rate)
(13) False Alarms = (Probability Knowing * 0) + (Probability Not Knowing * Default Rate)

Notice that these formulas assume that the child's knowledge of N+1, when it exists, has precisely the same properties as their knowledge of smaller numbers: a) it leads to an exact response when the number has been requested (the probability of a hit when knowledge is present is 1) and b) unknown numbers are interpreted as mutually exclusive with this term (the probability of a false alarm when knowledge is present is 0). By solving for these equations, we can determine both the likelihood that the child knew or retrieved the number on any given trial (the probability of knowing in (14)) and the likelihood of selecting N+1 when no knowledge is available (the default rate in (15)).

(14) Probability of Knowing = Hit Rate – False Alarm Rate
(15) Default Rate= (Hit Rate – Probability of Knowing) / (1 – Probability of Knowing)

As B&B note, the false alarm rate is not available for most studies which use the Give-N task because participants are typically tested on numbers of increasing magnitude until they fail, yielding little data for numbers N+2 or greater. One exception is Sarnecka and colleagues' study of American, Japanese and Russian children, in which all participants were tested on the numbers *one*, *two*, *three*, *five*, and *six* (Sarnecka et al., 2007). Following B&B, we used the raw data from this study to calculate the hit rate and false alarm rates for each child's highest known number (N) and for the next largest number (N+1). These values are given in Table 1 broken down by language group and knower level. For non-knowers, we cannot calculate values for N, since they do not know any numbers. Thus we focus our analysis on one-knowers and two-knowers.

INSERT TABLE 1 ABOUT HERE

As B&B note, the Sarnecka data demonstrate that children have some knowledge of the next highest word-quantity mapping: their hit rate is 62%, while their false alarm rate is 49%. However, this knowledge is not particularly robust. By the above formulas, children retrieve the meaning of N+1 on only 13% of the trials; the remainder of the time they simply exhibit a response bias to select N+1 for all number words greater than N (Default Rate = 56%). This

performance pattern contrasts sharply with their knowledge of N in this data set. Here the hit rate is 95%, while the false alarm rate is 2%, suggesting that children are retrieving the correct meaning on 93% of the trials.

These calculations assume that knowledge is discrete: on an individual trial it is either present or absent. If we wish to assume instead that knowledge (or knowledge retrieval) is graded, we can perform parallel analyses using the standard measures from signal detection theory. This analysis also suggests that knowledge of N+1 is quite weak (d-prime = .33, A-prime = .61) while knowledge of N is robust (d-prime = 3.78, A-prime = .98).

This pattern is not limited to the Give-N task. B&B provide parallel data for knowledge of N+1 from two studies in which children were asked to give numerical descriptions to sets of visible items (the What's-on-this-Card task, LeCorre & Carey, 2007). The composite of the two data sets paints a similar picture of children's partial knowledge (Hit Rate = 62%, False Alarm Rate = 49%, Probability of Knowing = 13%, Default Rate = 56%, d-prime = .33, A-prime = .62).[4]

These analyses demonstrate that the Give-N task does an impressive job of dividing children into discrete groups. Performance on the largest number that the child is deemed to know is near ceiling, whereas performance on the next largest number is above chance but quite poor. Moreover, the analyses yield estimates of partial knowledge that allow us to address B&B's arguments against partial knowledge (3.2.3) and to evaluate their explanation for the data from experiments on children's exact interpretation of number words (3.5).

3.2.3 Addressing B&B's arguments against partial knowledge

B&B offer two counter-arguments to the partial knowledge hypothesis. First, they suggest that the partial knowledge hypothesis cannot explain why children often give N+1 objects when they are asked for quantities N+2 or higher. Unknown numbers are typically interpreted as mutually exclusive with known numbers. Thus B&B conclude that partial knowledge of N+1 should prevent this quantity from being produced in response to other number words. But in actuality this pattern is not particularly problematic for the partial knowledge hypotheses that we

[4] If we examine the two samples separately the results are surprising. The children in the Le Corre and Carey (2007) study (N=48) show little knowledge of N+1 (hit rate=58%, false alarm rate = 54%, percentage known = 4%, d-prime = .10, A-prime = .54). In contrast those in the B&B data set (N=14) perform very well (hit rate =73%, false alarm rate = 31%, proportion known = 42%, d-prime = 1.01, A-prime = .80). This raises the possibility that children's knowledge of N+1 may vary across samples or populations. This is expected on most versions of the partial knowledge hypothesis; children's understanding of N+1 should depend on how much progress they have made in solidifying that mapping (and thus how often they are correctly retrieving it). On the face of it, B&B's theory should predict that all N-knowers will have similar and complete knowledge of N+1, since it is this knowledge that allows them to consistently assign an exact interpretation to N. However from the information available, we cannot determine whether the difference between the studies is reliable or rule out the possibility that it is driven by differences in the how the tasks were administered.

proposed. Because the knowledge is partial, it is not always available. When it is present it constrains interpretation, when it is absent it does not. This is built into the equations given above (the probability of giving N+1 for N+2 is 0 when the meaning of N+1 has been retrieved and equivalent to the default rate when it has not been retrieved). The theory that these equations specify is consistent with any data pattern other than one in which the false alarm rate is reliably greater than the hit rate, but such a pattern would also be unexpected on the B&B hypothesis. In fact, as we will see below (3.4.2), the non-exclusivity of N+1 is actually more problematic for B&B, forcing them to complicate their theory of how unknown numbers are interpreted.

Second, B&B also argue that the partial knowledge hypothesis is also inconsistent with findings from another experiment, Wynn's Point-to-X task (1992). In this study, children are asked to point to sets of different quantities. When one (or both) sets are within the child's known number range, performance is quite high. For our purposes the critical trials are those that contrast sets of N+1 items with sets of N+2 items. For example, trials in which one-knowers are shown a set of 2 and a set of 3 and asked for either *two* or *three*. Wynn found that children were at chance in this task, appearing to select the set at random. However, B&B note that if kids had partial, exact knowledge of *two* then this knowledge should allow them to succeed at least some of the time, pushing their performance above chance.

We accept this prediction but question whether the Wynn study had the sensitivity to detect such a difference. To ascertain this we would need to know: a) how much partial knowledge of N+1 would have to be attributed to the Wynn's children in the Give-N task on the exact hypothesis and b) how sensitive the Point-to-X task is relative to the Give-N task. Neither of these facts is available. While children's performance in the Point-to-X task is broken down by knower-level, no direct comparison is made between the sensitivity of the two tasks. In the version of the Give-N task which Wynn used, children were not generally asked for quantities of N+2 or more, thus the false alarm rate cannot be calculated.

Nevertheless, if we are willing to assume that different samples of one- and two-knowers exhibit similar levels of partial knowledge and have similar levels of sensitivity in different tasks, we can use the data from the English-speaking children in the Sarnecka study to begin to explore this question. To estimate the relative sensitivity of the two tasks we compared children's performance for their highest known number (N) in Sarnecka's Give-N data and Wynn's Point-to-X data. These comparisons favor the Give-N task but the difference appears to be negligible (the estimated "proportion known" for Give-N is 94% for one-knowers and 92% for two-knowers, while for Point-to-X it is 92% and 84% respectively).

To estimate the partial knowledge of N+1 in English speaking one- and two-knowers, we used the information in Table 1 and equations (12) and (13). We calculated that one-knowers retrieved the meaning of *two* about 10% of the time and two-knowers retrieved *three* about 15% of the time. These estimates were used to predict the percent correct in the Wynn forced choice task.[5] The results suggest we should expect quite small effects: if the tasks are equally good at

[5] Because equal numbers of each trial type ("Point to N+1" and "Point to N+2") were presented and the data was collapsed, the formulas above could be simplified to: Percent correct =

triggering partial knowledge, one-knowers should be correct just 55% of the time, and two-knowers just 58% of the time. In Wynn's sample, one-knowers (n=6) were correct on 54% of trials, and two-knowers (n=9) on 55% of the trials. These observed values are strikingly similar to the values predicted by the partial knowledge hypothesis. Given that the observed values were not reliably different from chance, they cannot be different from the (numerically closer) predicted values. Thus there is no evidence that Wynn's pointing task patterns differently from the other tasks, and consequently no reason to reject the hypothesis that some children, some of the time, have partial knowledge of the exact meaning of N+1 at the point when we label them N-knowers.[6]

3.3. Other arguments for lower boundedness

B&B offer three additional arguments in support of the claim that the meanings of number words are lower bounded, with exact interpretations arising via implicature.

3.3.1. Positing distinct meanings for "a" and "one" creates a learnability problem

In a truth-value judgment task, Barner and colleagues (2009) recently found that young children happily accept singular indefinites as descriptions of sets with 2 or more members, suggesting that *a* has a lower-bounded meaning. If *one* has an exact meaning, then children would have to learn distinct meanings for each term, presumably by observing differences in how the two terms were used. But in parallel tasks, adults restricted both *one* and *a* to sets with exactly 1 member. B&B suggest that children lack sufficient evidence to learn distinct meanings for these two words and thus the early differences in interpretation must arise from a common lower-bounded semantics and a systematic difference in the probability of implicature due to the relative salience of the scales.

However, like all poverty of the stimulus arguments, the validity of this one depends on a clear characterization of the information that is available in the input and the nature of the learning device that extracts it. Can we assume that input distinguishing *a* from *one* is limited? Intuition suggests that the distribution of two words differs substantially. In fact, there are few contexts in which it is natural to substitute one term for the other (see (16) and (17)).

(16) a. A man walked into a bar….

(Proportion Known * 1) + (Proportion Unknown *.5). Thus it was unnecessary (and impossible) to calculate the Default Rate of N+1 choices in this task.

[6] The above analysis assumed that knowledge is either present or absent. However, a similar picture emerges if we assume knowledge is graded. In this case calculating a measure of partial knowledge requires that we make assumptions about the distribution of the errors in the Wynn pointing task. Making the conservative assumption that errors are symmetric results in the lowest estimates of partial knowledge in this task (for one-knowers d-prime = .25, A-prime = .59; for two-knowers d-prime = .30, A-prime = .61). These values are similar to the estimates derived from Sarnecka's data for the English speakers in the Give-N task (for one-knowers d-prime = .28, A-prime = .60; for two-knowers d-prime = .49, A-prime = .64) and well within the range of values observed in the What's-on-this-Card task (see fn 5).

b. One man walked into one bar….

(17) a. More than one person has told me I'm obnoxious.
b. More than a person has told me I'm obnoxious.

Ultimately, far more work would be needed to characterize these differences and determine whether they are of use to the child, but it would be premature to assume that the information is not there. In fact B&B's account highlights a critical distributional difference between the forms: only *one* appears as part of the count list. B&B suggest that the count list increases the salience of the scale, allowing children to generate the implicatures for numbers long before they reliably generate implicatures for quantifiers. But the appearance of a word in such a list could also serve as a cue that this term has a meaning with both a lower and an upper bound. It indicates that *one* is a member of a semantic class with ten or more members (most one-knowers can recite the count list to *ten* or beyond). This may allow children to rule out other natural classes of quantificational concepts (e.g., the set consisting of NONE, A, SOME, ALL), and highlight the relevance of other classes (e.g., exact quantities, approximate numerical magnitudes, or mutually-exclusive parsings of the quantity scale).

3.3.2. Children's interpretation of unknown numbers uses the same abilities as scalar implicature, suggesting they have the necessary skills by two years of age.

Wynn (1992) noted that children treat *unknown* number words as if they were mutually exclusive with *known* number words. For example, when a two-knower is asked for *"five fish"*, she might give the experimenter 3 or 4, but would almost never give just 1 or 2. To do this, the child must recognize that numbers form a class and that sets that can be described with one term from that class cannot be described by other terms from this class. B&B point out that this process can be thought of as a kind of scalar implicature. Perhaps a two-knower maps an unknown number word such as *five* onto a broad numerical concept like AN UNSPECIFIED CARDINALITY. This meaning does not rule out the possibility of using *five* to refer to sets of 1 or 2. However, children are able to make the pragmatic inference that if you wanted them to give you 2 objects you would use the more specific term. On this analysis, the mutually exclusive interpretation of unknown number words requires an implicature, leading B&B to conclude that children can calculate implicatures for numbers and thus their difficulties with implicature may be limited to other scalar terms.

But processes that can be described in similar terms are not necessarily subserved by the same mechanisms. To explain the interpretation of unknown numbers, Wynn pursues an analysis based on Markman's description of how lexical constraints affect children's acquisition of object labels (1989). Markman proposed that children initially assume that words refer to whole objects (the whole-object constraint) and are extended to all members of the same basic-level kind (the taxonomic constraint). Thus children have a class of linguistic forms ("labels") that they are mapping to a class of concepts ("kinds"). This mapping, she argued, is constrained by mutual exclusivity: the child's assumption that each object has only one label.

This form of mutual exclusivity could reflect a simple computational preference for one-to-one mappings between concepts and forms. Such a mechanism would be very different from the kinds of mechanisms used to explain implicature. Take for example the two-knower. This child

has identified a set of forms (number words) and believes that they map onto a set of concepts. But she has only succeeded in making stable mappings for two of these labels (*one* and *two*). However, mutual exclusivity prevents her from mapping other forms in this class to the concepts linked to *one* and *two*. This kind of mutual exclusivity could be instantiated as strong stable connections between the known words and their conceptual representations and weak probabilistic connections between the remaining forms and concepts.[7]

On this theory, the process by which children avoid mapping *five* to TWO differs from scalar implicature in the following respects: a) the term being interpreted does not have a stable meaning which is overridden or supplemented; b) the terms do not form a scale based on their informational strength (*two* cannot be mapped to any of the candidate concepts for *five* just as *five* cannot be mapped to TWO); c) mutual exclusivity is not necessarily a post-semantic process, in fact exclusion could occur during word recognition (when *five* is probabilistically mapped to larger quantity concepts but not TWO). On this construal, the interpretation of unknown number words does not provide evidence of early competence in implicature; it is only as relevant (or irrelevant) as children's ability to infer that *dax* cannot mean CAT because *cat* means CAT.

3.3.3 The evidence to date suggests that children have little difficulty with implicature when the scale is made salient.

[7] It is unclear precisely what numerical concepts two- and three-year-old children are entertaining as potential meanings for unknown numbers. They know these unknown number words are also mutually exclusive with each other; when an array is labeled as *eight* they assume it cannot be labeled with *four* (Condry & Spelke, 2008). However they do not appear to understand that these terms refer to stable exact numerosities; if children are shown an array and told that it has *eight* items, and then half of these items are taken away, they continue to select it as an example of *eight* as often as an unlabelled array (Condry & Spelke, 2008). However, unlike B&B, we do not think that this finding undermines the hypothesis that mutual exclusivity for number words is a lexical phenomenon. One possibility is that children only know that unknown number meanings refer to quantity and that they are, at the semantic level, mutually exclusive and thus must possess both upper and lower bounds. This initial representation can be thought of as an ordered list of conceptual placeholders. Learning a number word consists both of linking a particular phonological form to the placeholder that is in the correct ordinal position and linking that placeholder to some value (or range of values) provided by another cognitive system which represents quantity. Once this link is made, the value associated with the known word constrains the interpretation of the remaining placeholders. Such a representation would allow the two-knower to infer that all unknown numbers refer to quantities greater than two and are mutually exclusive with one another. But when a particular unknown number is temporarily associated with a particular quantity (as in Condry & Spelke, 2008), the child cannot make a stable mapping both because she does not know the ordinal position of this number (children at this stage fail to use order in the count list to determine the relations between unknown numbers) and because she has yet to infer that all the placeholder concepts refer to single cardinalities. After all, there are many mutually exclusive parsings of the quantity scale that do not have this property (e.g., 1, 2, 3-10, 11, 12-15, 16-100).

B&B argue that children's failure to calculate many scalar implicatures derives from their inability to retrieve scalar alternatives. When the scale is made salient, this difficulty is alleviated. Thus it is plausible that in case of numbers, in which the scale is memorized, even two- and three-year-old children are able to consistently calculate implicatures and derive exact readings.

There are two limitations to this argument. First, the work that B&B highlight examines the factors that influence interpretation in children around five years of age. This is a period during which implicature calculation appears to change rapidly (see e.g., Foppollo et al., under review). Thus one might expect that many factors would influence performance at this age which would not have an effect at the younger ages when children are first learning number word meanings.

Second, the work to date falls short of showing that scale availability actually leads children to calculate upper bounds. Calculating scalar implicatures requires that a listener: a) retrieve the scalar alternatives; b) recognize that the stronger alternative would be more informative; and c) infer from the use of the weaker term that the speaker was not in a position to assert the stronger claim (either because it was false or because s/he lacked the necessary information). Step c is necessary for deriving the upper bound, but many prior tasks could be solved on the basis of informativity alone (step b). To illustrate this we focus in on two findings highlighted by B&B.

As we noted earlier, Papafragou and Musolino (2003) found that five-year-old children who heard statements with weak scalar terms in a context in which the stronger term applied, typically accepted those statements, suggesting that they did not compute the implicature. In contrast, adults typically rejected them. However, when children were given practice in detecting and correcting infelicitous statements, their performance improved and they began rejecting the statements about half the time. B&B argue that this demonstrates that children can calculate implicatures when the need to do so is made clear. But another interpretation of these findings is possible. Pre-training and other changes in the task may have led participants to shift the basis of their judgments from meaning (pragmatically enriched or not) to form. In the first experiment, participants were simply asked whether the puppet had answered well. While the criterion for making this judgment was left open, both the structure of the task (stories followed by descriptions), and the filler items (for the children mostly false statements) presumably focused participants on the meaning conveyed. In the second experiment, the participants' focus was shifted to the manner in which the statement was expressed: the stated goal of the task was to teach the puppet to speak better and the practice trials focused on violations of the maxim of manner (Grice, 1975). To extend this training to the scalar items, children would have had to recognize that the weaker terms were less informative, but we cannot tell whether they actually used this information to calculate an upper bound.

This same ambiguity is present in the second experiment that B&B discuss. Chierchia and colleagues (2001) found that five-year-olds fail to interpret *or* as excluding *and*—they are willing to accept a description like *"Every boy chose a skateboard or a bike"* in a context where some of the boys chose both a skateboard and a bike. However, these children do understand that in these contexts *and* is more felicitous than *or*: when given both descriptions they select the more informative one 90% of the time. B&B interpret this as evidence that children have the ability to calculate implicatures but generally fail to do so because they do not retrieve the scalar

alternatives (see also Gualmini et al., 2001; Reinhart, 1999). But again this study falls short of showing that children can calculate implicatures when the scale is provided. To successfully select the best utterance the child only needs to compare the informativeness of the two utterances (step b), she does not necessarily have to make the upper-bounding inference (step c).

Recent work by Foppollo and colleagues (under review) highlights this gap between informativeness and scalar implicature. In a series of experiments, they replicate the finding that five-year-olds often accept *some* in contexts where *all* applies, and again they find that the same children overwhelmingly succeed at the sentence selection task (preferring descriptions with *all*). However, the gap between these two tasks is more profound than B&B's theory of scale salience would predict. Priming the relevant scale has absolutely no effect on judgments: children who have just heard a parallel event described with *all* are no more likely to reject an underinformative utterance with *some*.

3.4. Theory-internal problems and problematic resolutions

Two facets of the data from the Give-N task are puzzling on B&B's theory.

3.4.1 Why are children willing to allow unknown number words to refer to sets of N+1?

The Sarnecka data indicate that on 49% of the trials in which children are asked for N+2 or more they hand the experimenter exactly N+1 items. This behavior is potentially problematic for B&B's account. As we noted earlier, children appear to interpret unknown numbers as contrasting in meaning with known numbers. Taken at face value, this finding suggests that if the child already knows that "N+1" means at least N+1, as B&B propose, then she should be able to infer that "N+2" cannot mean N+1 and thus must refer to a larger number.[8]

One might be tempted to explain these responses as errors: perhaps some of the children do not know the meaning of N+1 or fail to retrieve it on some of the trials and thus do not interpret larger numbers as excluding this quantity. This is indeed our explanation, but this option is not open to B&B. By their hypothesis, N-knowers treat N as exact only because they know the meaning of N+1 and are able to retrieve it reliably. Since such children give exact responses to N over 90% of the time, they must retrieve N+1 successfully on at least 90% of the trials so as to perform the implicature that restricts the application of N to exactly N objects.

B&B are aware of this problem, and they address it by positing a distinction, between how implicatures are calculated for known and unknown numbers. For known number words, the scalar alternatives are the *lower-bounded* meanings of known number words, just as in previous theories. But for unknown number words, they argue that the scalar alternatives are the upper-bounded interpretations which have been derived via scalar implicature. Because N-knowers, by hypothesis, have no known number word that strengthens N+1, B&B argue that they cannot use their knowledge of the meaning of N+1 to restrict their application of unknown number words such as N+2.

---

[8] As we noted above, this behavior is readily explained by our partial knowledge hypothesis (these are the trials on which the child did not have access to the meaning of N+1).

B&B's solution is problematic for two reasons. First, it introduces an asymmetry between implicature for known vs. unknown number words: scalar implicature is necessary to make something an alternative if the term in question is unknown, but not if it is known. Take for example a child who appears to be one-knower. By hypothesis, this child knows that *two* means AT LEAST TWO. Upon encountering *one,* the child retrieves the meaning of *two,* which is lower-bounded and uses it to calculate the upper bound of *one*. In contrast, to infer the lower bound of an unknown word the child must first calculate implicatures for her known number words. Thus when the one-knower is asked for *three*, she first retrieves the alternative known words (*one* and *two*), calculates the relevant implicature to determine which have strengthened meanings (*one* is EXACTLY ONE), and then rules out these cardinalities as referents for the term in question (*three* is not ONE). Positing distinct processes for known and unknown numbers undermines previous arguments that the machinery for early implicatures is evident in toddlers' interpretation of unknown numbers. Furthermore, it raises questions about how children would discover which implicature processes operate on meanings that have already been strengthened by previous implicatures and which do not.

Second, this wrinkle in the theory drastically increases the computational burden on the child. Previous studies demonstrate that five-year-old children regularly fail to calculate implicatures for a variety of non-numerical scales. B&B's original theory posited that the salience of the count list allows two- and three-year-olds to overcome this limitation and calculate a single upper bound for known numbers. Their augmented theory then credits children with the ability to reliably make a handful of first-order implicatures followed by a second-order one. For example, to infer that *five* is not ONE, TWO or THREE, the three-knower must retrieve four terms (*one*, *two*, *three* and *four*), calculate three first-order implicatures (EXACTLY ONE, EXACTLY TWO, EXACTLY THREE) and then use these as input into a second-order implicature.

To date there is no strong evidence that three-year-olds *ever* calculate non-numerical implicatures, but their interpretation of known number words is unswervingly exact. If number words have exact meanings, this discrepancy is readily accounted for. If they do not, our theories are rapidly forced to greater complexity.

3.4.2 Why do children give N+1 preferentially for N+1?

If a word has a lower-bounded meaning and there is no stronger term to create an implicature, then all quantities that are consistent with this lower bound are equally good instances of this quantity. Thus on B&B's theory, it is unclear why children would prefer to pair N+1 with sets of exactly N+1 (rather than sets of N+2 or more). To explore the basis of this bias, B&B conducted a study in which adults are asked for *"at least X*" (for X = 1, 2, 3). By hypothesis *"at least X"* should have the same meaning for adults as N+1 does for the N-Knower. They find that adults share this preference, selecting exactly X on about half the trials. Thus they argue that it stands to reason that children who interpret N+1 as lower-bounded should do the same.

This fact is interesting but difficult to interpret in the absence of a theory about *why* adults show this behavior. One possibility is that adults interpret an utterance like *"at least two"* as contrasting with utterances like *"at least three"* or *"at least four."* Such an interpretation seems

particularly likely in this experimental context in which the different numbers were presented to the same subjects and no clear motivation was given for the use of such precise phrasing for such (semantically) vague requests. On this account, the adults' preference for giving exactly X depends on their knowledge of the phrase that encodes X+1. But this is precisely the knowledge that B&B argue the N-knower lacks.

3.5. The box task revisited

These two hypotheses about children's knowledge of N+1 have very different implications for our interpretation of performance in the number conditions in the box task. On our partial knowledge hypothesis, our interpretation of this data depends largely on the degree of partial knowledge that two-knowers are likely to have for *three*. Our estimates suggest this knowledge is quite weak. For example, we calculated that children access the meaning of N+1 only about 13% of the time in the Give-N task. Our analyses of the What's-on-this-Card and Point-to-X tasks suggested that they elicit a similar degree of partial knowledge. It is unlikely that the box task would elicit partial knowledge to a greater degree (particularly since the term for the larger quantity is never mentioned). Thus our best guess is that about 87% of the time, the two-knowers in our experiment did not access (or did not know) the meaning of *three*. Thus if knowledge of *three* was necessary to calculate the upper boundary of *two*, children should have selected the box with 3 fish on the vast majority of the trials. Instead they categorically rejected it and inferred that covered box must contain 2 fish.

In contrast, on B&B's lower-bounded hypothesis, the children that we tested in the box task were not two-knowers but three-knowers: they interpreted *two* exactly, not because it has an exact meaning, but because they already understood that *three* means AT LEAST THREE, and were able to use this term to infer that *two* could not mean THREE. For this hypothesis, the greatest challenge is to explain why this exact interpretation persisted in a context where the scalar implicature for *some* was cancelled, even in adults. To meet this challenge, B&B could draw on the distinction that they make between salient scales (numbers) and less salient scales (other scalar terms). For example, one might argue that on the critical trials of the box task, the relevant alternatives are less salient to all participants because only one is visible in the referential context. In the case of the quantifier scale, this prevents even adults from accessing *all* as an alternative for *some*. However, the number scale is so salient that even young two-knowers, who by B&B's hypothesis have recently acquired *three*, are able to access *three* to provide an upper bound for *two*.[9]

We see several reasons for preferring the partial knowledge account. First, it is consistent with what we know about the nature of development (Siegler, 2007) and the imperfection of psychological measurement (section 3.2.1). The notion that knower-levels represent discrete stages which can be perfectly ascertained by a single task is an idealization that has been useful but is bound to be false in the limit. Second, the partial knowledge account does not require that

[9] One challenge for such an account is finding a noncircular way to determine which contexts make scalar alternatives less salient. One might think that the contrast between *all* and *some* would be very salient on the critical trials of the box task: participants see a situation consistent with *all* and hear a description using *some*.

we posit additional mechanisms to explain why children prefer to give N+1 for N+1 or why they fail to treat other unknown numbers as mutually exclusive with this term (section 3.4). Third, the B&B hypothesis requires that we assume that two- and three-year-olds calculate implicatures for numbers virtually 100% of the time. This is remarkable considering that all other scalar terms implicatures are often sluggish and variable well into middle childhood. This reasoning becomes even more problematic when we consider B&B's argument that children's interpretation of unknown numbers involves multiple first order implicatures and a second order one (3.4.1). Finally, we contend that the partial knowledge account, which posits exact meanings for numbers, provides the simplest explanation for the results of the box task. In a context in which implicatures are cancelled for scalar terms both adults and children insist upon exact readings for numbers, going out of their way to locate a match.

3.6 The difference is in the process

Nevertheless, we recognize that the data that are currently on the table can be squeezed into either theory. Both theories predict that numbers will generally be interpreted as exact, even in contexts and in populations where other scalar inferences fall through. Both theories offer an explanation for the children's above chance performance on N+1. In fact, the two theories do not appear to differ at all in their predictions about the interpretation that people assign to numbers. Critically, however, the two theories posit different processes to account for these interpretations. On B&B's account, an exact interpretation involves three steps: a lower-bounded meaning is semantically composed, the relevant alternatives on the scale are computed, and the inference is calculated. This theory makes clear predictions about the time course of comprehension: the lower-bounded semantic meaning should be visible at some point during processing prior to the time at which the scalar implicature is calculated. In contrast, on our account, the meanings of number words are exact from the moment they are encountered. Consequently, we should expect the upper bounds of number words to influence interpretation as rapidly and to the same extent as their semantically-encoded lower bounds.

Similar predictions motivate studies of adult language comprehension which use reaction time measures to probe the on-line processing of scalar terms (Rips, 1975; Noveck & Posada, 2003; De Neys & Schaeken, 2007). For example, Bott and Noveck (2004) compared response times for truth-value judgments of underinformative sentences like *"Some elephants are mammals*." They found that participants who spontaneously adopted an upper-bounded interpretation (judged the statement to be false) took longer than participants who adopted a lower-bounded interpretation (judged the statement to be true). Recently, we explored the nature of this delay by presenting adults and five-year-old children with commands like *"Point to the girl that has some of the pills"* and monitoring their eye-movements to displays featuring a girl with 2 of 4 pills and another with 3 of 3 pillows (Huang & Snedeker, 2009; Huang & Snedeker, in press). In the Huang and Snedeker studies, there is a critical period of ambiguity at the onset of the quantifier where the semantics of *some* is compatible with both characters. The calculation of the scalar implicature would eliminate this ambiguity restricting reference to the girl with the subset of items. We found that participants were substantially delayed in interpreting sentences that required this implicature, suggesting a temporal lag between semantic processing and the calculation of the pragmatic inference. Furthermore, while adults eventually

calculated the implicature (Huang & Snedeker 2009), we found no evidence that the children ever did (Huang & Snedeker, in press).

In addition to the critical *some* and *all* trials, these experiments also included trials where characters were referred to using number words (e.g., "*the girl that has two of the pills"*). The data on these trials provide an opportunity to probe the processes underlying adults' and children's interpretation of number words. If number words have lower-bounded meanings and the upper bound is calculated by scalar implicature, then participants should quickly use a number word to exclude a smaller set (rapidly inferring that *three* cannot refer to 2 objects) but should show delays in using a number word to exclude a larger set (failing immediately to infer that *two* cannot be 3). In contrast, if numbers have exact meanings, then both the upper bound and the lower bound should be accessible as soon as the meaning is accessed, resulting in equivalently fast reference resolution in both directions.

Furthermore, we might expect this pattern to be more exaggerated in children than in adults, for several reasons. First, if educated adults acquire exact meanings for the mathematical numerals through formal education, as Levinson (2000) argues, then this competing interpretation should be unavailable to children who have had little formal instruction. Second, prior studies have found that scalar implicatures are cognitively taxing even for adults who generate these inferences frequently and reliably (Noveck & Posada, 2003; Bott & Noveck, 2004; De Neys & Schaeken, 2007). The cost of implicature is likely to be even greater in young children, who have smaller memory spans and process information more slowly (Dempster, 1981; Schneider & Bjorklund, 1998; Kail, 1991; Kail & Salthouse, 1994). These limitations in cognitive processing may hinder children's ability to retrieve the stronger alternative during real-time comprehension leading to delays in generating an upper bound (Gualmini et al., 2001; Reinhart, 1999; Pouscoulous et al., 2007).

4. A reanalysis of the data from Huang & Snedeker (2009/in press)

Below we revisit two published experiments on the comprehension of numbers and quantifiers: one study with adults (Experiment 2, Huang & Snedeker, 2009) and a tightly matched study on five-year-old children (Experiment 1, Huang & Snedeker, in press). The prior papers focused solely on the processes underlying adults' and children's interpretation of the scalar quantifiers, and their implications for a theory of implicature. Here we turn our attention to adults' and children's processing of the number words.

4.1. Methods

Twenty undergraduates and 24 five-year-old children were presented with stories accompanied by a visual display in which two types of objects were divided up between two boys and two girls (see Figure 5). These objects were labeled by nouns whose initial phonemes were identical, On the *two* and *three* trials, these items were divided such that one of the critical characters (e.g., one of the girls) had 2 out of 4 of one type of object (e.g., pills) and another had 3 out of 4 of the other type of object (e.g. pillows). In the *some* and *all* trials, the critical

characters had either 2 out of 4 of one type of object or 3 out of 3 of the other.[10] Immediately following each story, participants' eye-movements were measured to the display as they heard commands like (18) below.

(18) Point to the girl that has <u>two/three/some/all</u> of the pills.

INSERT FIGURE 5 ABOUT HERE

Critically, this design allows us to investigate the meaning of number words by examining the pattern of eye-movements in the *two* trials. If number words have lower-bounded semantics, then we should expect a delay in reference resolution, as occurs for the pragmatically-bounded term *some*. If, however, numbers have exact semantics, then we should expect reference resolution to be as rapid for *two* as it is for *three,* where the meaning of the number word immediately rules out the same gender Distractor (the girl with 2 pillows).

4.2. Results

In our primary analyses, the dependent measure was looking time to the Target as a proportion of looking time to the Target and the Distractor. This score ranged from zero (exclusive looks to the Distractor) to one (exclusive looks to the Target). Looks to the other characters were infrequent after onset of the gender cue (less than 5% in every condition) and were not included in the analysis.

Adult data. Figure 6 illustrates the fixation patterns for adults in each of the four conditions across different time windows corresponding to the instructions. Prior to the onset of the quantifier, there were no systematic differences in Target looks across the conditions (all p's > .50). In the region immediately following the onset of the quantifier, adults in both the *all* and *three* trials showed a reliable preference for the Target (p's < .05), demonstrating that the semantics of both terms could be used to rapidly resolve the referent of the phrase. In the *some* trials however, there was no reliable preference for the Target ($p > .50$). This pattern demonstrates that our task was sensitive to the temporal ordering of semantic and pragmatic processes: during this ambiguous period, the participants accessed the lower-bounded semantics of *some* but failed to make the implicature which would rule out the Distractor as a potential referent.

INSERT FIGURE 6 ABOUT HERE

Crucially during this ambiguous period, there was also a reliable Target preference for *two* trials ($p < .05$): Participants strongly preferred sets of exactly two, over sets of exactly three. Reference resolution was as strong for *two* as it was for *all* and *three* (both p's > .50), suggesting that the disambiguating upper bound of *two*, like the lower bound of *three* or *all*, is retrieved with

[10] This difference between the number and quantifier trials was necessary to ensure that the verbal descriptions were felicitous for all trial types. While saying "*three of the pills*" would be odd when there are only 3 pills in total, adding an extra object to the character of opposite gender makes the utterance felicitous without changing the visual properties of the critical Target or Distractor characters (Huang & Snedeker, 2009).

the lexical item. In contrast, the Target preference for *two* was reliably greater than the Target preference for *some* ($p < .01$), highlighting the difference between this number word and a true scalar term. Finally, following the onset of phonological disambiguation, adults responded rapidly by shifting looks to the Target in all four trial types.

To further clarify how eye-movements changed in response to the quantifier, we separated the trials based on whether participants were or were not fixating on the Target during the frame immediately preceding the onset of the quantifier (Swingley & Fernald, 2002). On trials where participants were initially fixating on the Distractor, we calculated the probability that they had switched their gaze to the Target for time window following the quantifier. On trials where they were initially fixating on the Target, we calculated the probability that they abandoned the correct referent in favor of the Distractor. Both the Lower-bounded and Exact accounts predict there should be rapid switching to the Target following *three* and *all* but fewer switches to the Target and more switches off the Target following *some*. However, for *two*, the Lower-bounded hypothesis predicts greater switches off the Target while the Exact account predicts greater switches onto the Target.

We performed these analyses on 100 ms intervals beginning from the onset of the quantifier (Figure 7). During the Distractor initial trials, the pattern of looks across the four terms differed from the 600 ms to 800 ms time window. In the *all* and *three* trials, adults began to switch their looks to the Target while in the *some* trials, they were more likely to continue looking at the Distractor. Critically, switches to the Target in the *two* trials grouped with the *three* and *all* trials ($p > .50$) and were significantly higher than the *some* trials ($p < .01$). A similar pattern emerged in the Target initial trials. During the 200 ms time window, adults were more likely to abandon their early looks to the Target after hearing *some* relative to *three* and *all* (p's $< .05$). In contrast, switches off the Target in the *two* trials were again no different than those for the *three* and *all* trials ($p > .50$) but were reliably less than the switches off the Target for *some* ($p < .05$).

INSERT FIGURE 7 ABOUT HERE

Altogether, these results suggest that the upper bound of a number term is available as rapidly as the lower bound for adults. *Two* excludes a set of *three* as rapidly as *three* excludes a set of *two*. In contrast, the pragmatically inferred upper bound of the scalar quantifier *some* was not immediately available. While adults were quickly able to rule out a subset after encountering *all*, they continued to consider a total set as a potential referent for *some* until the utterance was phonologically disambiguated. These results demonstrate that adults interpret number words differently from scalars and that they access an exact meaning for numbers at the earliest moments of language processing. It is possible, however, that adults' bias to interpret number words exactly derives from their lifetime of experience with exact numbers in economics (e.g., prices) and arithmetic. We turn, therefore, to the findings with children who have little or no experience with money or formal mathematics.

Children's data. Figure 8 illustrates that the children's pattern of reference resolution broadly paralleled that of the adults. Like adults, there was a period after the onset of the quantifier where five-year-olds exhibited a Target preference for *all*, *three*, and *two* but not *some*. And like the adults, they eventually converged on the Target in all four conditions. Unlike

adults, however, children showed preferences for particular quantities which emerged prior to the quantifier onset, complicating interpretation of this analysis. Thus we again separated trials based on where the participant was looking immediately prior to the onset of the quantifier. By comparing trials on which children were looking at the same objects when the quantifier began we can factor out differences in fixation patterns that may be driven by perceptual biases, and focus on effects which emerge in response to these critical terms.

INSERT FIGURE 8 ABOUT HERE

On the Distractor initial trials, the pattern of switches began to differ across the four terms about 800 ms after the onset of the quantifiers (Figure 9). As expected, during the *all* and *three* trials, children began switching to the Target item, presumably because the semantics of these terms ruled out the Distractor. In contrast during the *some* trials, children were more likely to continue looking at the Distractor, suggesting that they had failed to calculate the upward bounding inference. Critically, switches to the Target during the *two* trials were no different than switches during the *all* and *three* trials (all p's > .50), indicating that the upper bound of *two* is also rapidly encoded. Moreover, switches to the Target were significantly lower during the *some* trials than during the *all, three,* and *two* trials (all p's < .01). This pattern across the trial types were even more pronounced in the Target initial trials. During the 500 ms to 700 ms window, children were more likely to abandon their initial looks to the Target after hearing *some* compared to *three* and *all* ($p < .05$). This suggests that for children, like adults, the scalar implicature was not initially available to resolve the referent of *some*. Critically, switches off the Target in the *two* trials were no different than those for the *three* and *all* trials ($p > .50$), suggesting that children had access to the upper bound for this number word as early as they had access to the lower-bound of *three* and *all*. However, these switches for *two* were reliably less frequent than those for *some* ($p < .05$).

INSERT FIGURE 9 ABOUT HERE

4.3. Discussion

The results from this experiment highlight two distinct patterns of reference resolution among the quantifiers. First, we found that reference resolution via semantics is rapid in both adults and children. Following the onset of *three* and *all*, both groups of participants were able to quickly rule out the semantically incongruent character and exhibit an early preference for the correct referent. Second, we found that reference resolution was delayed when it required the generation of a pragmatically-specified upper bound. Following the onset of *some*, both adults and children continued to look at both characters, suggesting the scalar implicature was not available immediately. The presence of these two patterns provides an opportunity to tease apart the Lower-bounded and Exact accounts. Our results from the *two* trials demonstrate that the upper bound for this term was available as quickly as the semantically-specified lower bounds for *three* and *all*. This led to an early preference for the correct referent in adults and a greater likelihood to adhere to this target in children. These findings provide strong evidence that number word meanings are exact: from the earliest moments of interpretation *two* has an upper bound while *some*, a true scalar term, does not.

4.4. Can the Lower-bounded hypothesis be saved?

Earlier we noted that theories positing a lower-bounded semantics for numbers come in two kinds. While some authors propose that the number *words* have lower-bounded meanings (e.g., Horn, 1972 & 1989; Gadzar, 1979; Levinson, 2000; Winter, 2001), others suggest that the lower-bounded meaning of numerically quantified noun phrases arises as a result of compositional processes (e.g., Fox & Hackl, 2004; van Rooy & Shulz, 2006; Ionin & Matushansky, 2006; Chierchia, Fox, & Spector, 2008; Barner & Bachrach, in press; Panizza et al, in press). We have framed our discussion in terms of the meanings of number words, thus one might wonder to what degree it applies to the compositional hypothesis.

Many researchers advocating the compositional approach have made predictions parallel to those tested here. For example, Barner and Bachrach (in press) clearly predict that a lower-bounded semantics should result in willingness to accept larger quantities if the implicature is unavailable (because the child does not know the more informative term). Similarly Panizza and colleagues (in press) argue that calculating the upper bound of a number is a costly operation which often results in a measurable delay. Even those who make no clear psycholinguistic predictions, typically argue that the same exact processes are employed in determining the upper bound of numbers and scalar quantifiers, leading to the prediction that the two classes should show parallel signatures in processing and in development. Our findings undermine these claims. In both of our studies, number words consistently failed to show the patterns observed for the scalar quantifier *some*. Thus resurrection of the lower-bounded hypothesis requires that it be cut free from these predictions. To do so, these theories would have to posit that these differences in performance between numbers and scalar quantifiers are attributable to a level of representation that is either higher or lower than the compositional level at which they are argued to have a parallel representation.

For example, one might argue that the differences in interpretation for *some* and *two* in our processing studies reflect a difference in the lexical representation of these terms (see Pannizza et al., in press). For example, if the lexical meaning of *two* is exact (Ionin & Matushansky, 2006), then participants may be able to use this meaning to perceptually pick out sets consisting of two physical objects before compositional processes kick in, linking the number and the noun, and removing the upper bound. However, this account does not provide an elegant explanation for the data from the box task. In the critical trials, there is no visible set of 2 objects that could be perceptually linked to this lexical representation prior to full semantic composition. Thus this theory would need to be supplemented with hypotheses about difference in the robustness of implicature for different classes of scalar terms (see discussion of B&B above).

A theory which attributes these asymmetries to a higher level of representation might provide a more natural account of the data. On theories in which the at least interpretation arises solely through a compositional process of existential closure, a statement like (19) has a semantic interpretation that can be glossed roughly and casually as (20).

19, John has two fish.
20. There exist two things such that they are fish and John has them.

Semanticists rightly note that a statement like (20) in no way rules out a statement like (21).

21. There exist three things such that they are fish and John has them.

Thus they conclude that *"two fish"* means at least two fish, since it is logically possible that there are more. But the processing mechanisms that mediate between semantic forms and nonlinguistic representations may fail to pass information about this logical possibility. Perhaps when we interpret a semantic representation like (20), we update our discourse model with a representation in which John has exactly two fish, no more and no less.[11] This representations does not explicitly deny (21), thus no implicature has been made. In fact, when pushed into a corner, we might admit to the possibility of (21), but before it was mentioned, we may never have considered it.

This account appears to be compatible with the experimental results to date. This higher level representation of an exact quantity (exactly two fish) could guide rapid reference resolution in the eye-tracking studies and lead children to provide exactly 2 objects when *two* are requested and could lead both children and adults to form a clear expectation for what should be in the requested box. Intuitively, such an account appears to make quite different predictions for *some*. A statement with *some* like (22) is typically argued to have a semantic interpretation with the rough gloss given in (23).

22, John has some of the fish.
23. For some things such that they are fish (in a contextually relevant set), John has them.

When we update our discourse model on the basis of (23), we may simply enter a representation of John possessing a quantity of fish, The status of any fish that John does not have is not explicit in (23), thus our model might initially lack any representation of the complement set. Such a representation would be consistent with both the Target and Distractor in the eye-tracking study, accounting for the delayed resolution of the referent, and would be consistent with the box containing all the fish on the critical trials of the box task. On this account, ruling out the total set would require the process of scalar implicature.

Notice that on this account the interpretation of number words is very similar to other terms that are not typically considered to be scalar. Take for example the sentence in (24). Logically it is compatible with (25)

24. John has a dog.
25. John has a dog and cat.

In fact, we could readily construct a context in which a dog and a cat are seen as contextually relevant alternatives and the utterance of (24) is taken to imply that (25) is not true. Thus the relation between "a dog" and "a dog and a cat" is much like the relation between "two fish" and "three fish". But it is unlikely that every utterance of "a dog" leads us to develop a discourse model that includes a dog and the possibility of a cat (a rat, a bad case of bronchitis, etc). Instead

[11] While we describe this possibility in terms of the representations in the discourse model, these effects could arise at any level of representation subsequent to the semantic representation and common to the various tasks under discussion.

we presumably just represent the dog (the exact meaning) but only explicitly enter the absence of a cat when the context strongly support this inference.

5. General Discussion

5.1. Findings

This paper sought to use experimental data to tease apart the semantic and pragmatic contributions to the interpretation of number words. First, we presented data on number word interpretation during early acquisition. Consistent with the Exact semantics account, we found that two- and three-year-olds interpreted words like *two* as referring to an exact quantity at the earliest stage of development. Like adults, they were able to reject salient lower-bounded targets and to use the number word to infer the presence of an exact match elsewhere in the array (i.e. in the covered box). Their interpretation of number words contrasts with their interpretation of true scalar terms like *some,* as both children and adults readily selected ALL as an example of *some* when there was no other visible alternative.

Next, we addressed a recent proposal that the acquisition pattern for number words suggests that they have a lower-bounded semantics. We argue that the evidence for a lower-bounded meaning is in fact more compatible with a theory in which children who have acquired a given number (N) sometimes have partial knowledge of the next word-to-number mapping (N+1). When this knowledge is present, children interpret the word as exact, when it is absent the unknown number is interpreted just like other unknown numbers (it contrasts with smaller numbers but is indistinguishable from other unknown numbers).

Finally, we re-examined data from studies of on-line language comprehension to compare the interpretation of number words with the interpretation of scalar quantifiers. Consistent with the Exact semantics account, adults and five-year-old children generated the upper bound for *two* as quickly as the semantically-specified lower bounds of *three* and *all*. In contrast, when presented with a lower-bounded quantifier like *some*, children, like adults, initially failed to generate an implicature that would rule out the total set. This finding provides evidence that unlike true scalars, the upper bound for number words is semantically encoded.

5.2. Methodological contributions

The present research makes two contributions to the empirical study of semantics, pragmatics, and their interface. First, it introduces the covered box task as a method for mapping the semantic boundaries of words and phrases. Tasks in which the meanings of words and expressions are assessed by presenting participants with a choice among potential referents create a strong demand on participants to select the best item from the set of options presented. Typically these studies focus on the status of marginal category members or pragmatically infelicitous interpretations. For such questions the tasks demands of typical choice paradigms can create one of two problems. First, if the questionable category members are contrasted with typical category members (*"Give me the birds"* in a context with sparrows and penguins) the semantic meaning of the term may appear to be narrower than it actually is. In a choice task, participants may prefer typical exemplars and fail to select less typical exemplars, even if both fall within the extension of the tested word or phrase, because they construe their task as choosing the best exemplars rather than all possible exemplars. This problem may be magnified

in young children, who face limitations of memory and attention and thus may be more likely to get distracted after making an initial selection. However, if uncontroversial category members are not included in the set of alternatives, a second and equally serious problem arises. The demands of the task may lead subjects to stretch the meanings of words and phrases in ways that do not reflect their true meanings (see for example Syrett, Kennedy, & Lidz, in press). If no true referents of a word or expression are presented (*"Give me the fish"* in a context with dogs and whales) participants may redefine their task as choosing the alternative that is "most like" the request and thus select a referent that is outside the true extension of the expression.

The box task avoids this demand by providing a foil which participants can interpret however they like. Our experiments demonstrate that when participants are given a definite description and are allowed to select an option that they cannot see, they will chose to do so only when none of the visible options match the meaning of the description. Thus the covered box task allows experimenters to test the extension of a description without making any a priori assumptions about the status of atypical exemplars (see Li, Barner, & Huang, 2008 for a further application of this method). Notably we found that even two-year-olds made systematic inferences about the contents of the covered box, suggesting that this task is appropriate across a wide age range.

The second methodological contribution of the present research comes from our use of the eye-tracking paradigm to measure processes of pragmatic interpretation. Methodologically, the eye-tracking paradigm has several advantages for exploring the relationship between semantics and pragmatics across development. First, since eye-movements are typically made without conscious reflection, they provide an implicit measure of comprehension which is well suited for studying populations with limited metalinguistic awareness. Second, 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. Third, because eye movements are well controlled by early childhood, the method can be used over a wide range of ages, from preschool children to adults (see Trueswell, Sekerina, Hill & Logrip, 1999; Snedeker & Trueswell, 2004). In the present case, the findings from our studies using this paradigm provide evidence that the upper bounds of number words are computed as quickly as their lower bounds, not only by adults but also by children with no history of arithmetic instruction. The failure of these methods to isolate a process of scalar implicature for number words, coupled with their successful isolation of this process in the case of the quantifier *some*, provides clear evidence for exact number word meanings.

Although number is our test case, we believe that both these methods can be usefully harnessed to explore the meanings of other words and expressions, and the ways in which these meaning are enriched by pragmatic inferences during language comprehension. These two methods have complementary strengths. The box task is easily employed, provides strikingly consistent response patterns, and measures the participants' ultimate interpretation. In contrast eye-tracking taps a behavior which is rapid and relatively automatic. Where the findings of the two methods converge, as in the present case, they provide robust evidence about the boundary between semantic meaning and pragmatic inference. Together, these methods may provide new empirical entry points into variety of old debates about the boundary of meaning and inference, allowing us to explore these processes across the lifespan.

5.3. Implications

Finally, the present findings speak to more general questions concerning the development of language and concepts. Unlike the two- and three-year-olds in our box task, adults have mastered the meanings of all the words in their count list. Through their experiences with money, measurement, and mathematics, they have come to use these words to express ideas far beyond that of young children. Yet with respect to language development, this research demonstrates a strong continuity between the semantic and interpretive processes of adults and young children. Adults and children both endow the number words in their lexicon with exact meanings, and appear to use the same processes to interpret these terms and apply them to sets of entities. Given the vast conceptual changes that occur in the domain of number during childhood (Carey, 2009) this developmental continuity in number word semantics is striking.

Concerning conceptual development, the present findings help to reconcile two large literatures on number words and number concepts. As we have discussed, the predominant linguistic theory of number semantics posits a large gap between the basic meanings of numerically quantified noun phrases (which is lower bounded) and the most common use of these phrases (to designate set with exact numerical quantities). As Levinson notes (2000), these lower-bounded meanings clearly could not support exact mathematical reasoning: Statements like "3 + 4 = 7" make sense only if all three numbers are interpreted as being strictly exact. As a consequence, he argues that number words are homophonous: Children initially acquire the lower-bounded number words of ordinary conversation, and then must later acquire through formal instruction, a new set of exact terms for mathematics and measurement. On this hypothesis, there is little connection between our informal and formal number concepts.

On the other hand, research in cognitive development provides a wealth of evidence that children's mastery of number word meanings directly supports their entry into formal mathematics. Children who have mastered the language of verbal counting perform far better in the kindergarten mathematics curriculum than those who do not, and interventions to enhance their number word mastery lead to improvements in their school mathematics achievement (Case & Griffin, 1996; Siegler, 2007). Moreover, children who have learned to count, but have not yet been taught any arithmetic in school, are spontaneously able to use counting to solve small-number addition problems exactly (Case & Griffin, 1996; Zur & Gelman, 2004) and large-number addition problems approximately (Gilmore & Spelke, 2008). Such achievements would seem to be impossible if the number word meanings that children initially mastered were lower-bounded, or if the number words in ordinary conversations and those in mathematical formalizations were learned as separate lexical entries.

5.4. Final Words

The present studies provide three good reasons to believe that number words are not bounded by scalar implicature, but instead have exact meanings. First, adults interpret numbers as exact even in a context in which scalar implicatures are cancelled. Second, young children who have assigned a meaning to *two*, but not *three*, interpret *two* exactly, indicating that mastery of the scalar alternative is not necessary for exact interpretation. Finally, both five-year-old children and adults rapidly use the upper bound of a number to infer the referent of a quantified phrase. In contrast adults are sluggish in calculating the upper boundary of a true scalar quantifier like

some and children simply fail to make this inference. With some squeezing, the first two pieces of evidence can be accommodated by Barner and Bachrach's theory of the development of lower-bounded meanings (in press). But we noted that this proposal has several drawbacks, both theoretical and empirical, and it cannot readily explain the results of the online processing studies.

In contrast, all of the existing evidence is compatible with the Exact semantics hypothesis, provided we assume: 1) that acquisition is not instantaneous and thus children sometimes have partial knowledge of a number word and 2) that there are pragmatic mechanisms for generating readings that initially appear to be lower-bounded (Breheney, 2008; Panizza, Chierchia, Huang, & Snedeker, 2009). Thus we conclude, that in this instance at least, cognitive reality matches our pretheoretical intuitions—numbers mean exactly what they appear to mean.

References

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.

Barner, B. & Bachrach, A. (in press). To know exactly one, children acquire at least two. To appear in *Cognitive Psychology*.

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

Breheny, R. (2008). A new look at the semantics and pragmatics of numerically quantified noun phrases. *Journal of Semantics*, 25, 93-139

Bultinck, B. (2005). Why Paul Grice should have been a corpus linguist: An analysis of two. *Paper presented at the 9th International Pragmatics Conference* at Riva del Garda, Italy.

Carey, S. (2009). *The Origin of Concepts.* New York: Oxford University Press.

Carston, R. (1998). Informativeness, relevance and scalar implicature. In: R. Carston & S. Uchida (eds.). *Relevance Theory: Applications and Implications*, 179-236. Amsterdam: John Benjamins.

Chierchia, G., Crain, S., Guasti, M. T., Gualmini, A., & Meroni, L. (2001). The acquisition of disjunction: Evidence for a grammatical view of scalar implicatures. In A. H.-J. Do, L. Domingues, & A. Johansen, (Eds.), *Proceedings of the 25th Boston University Conference on Language Development* (pp. 157-168). Somerville, MA: Cascadilla Press.

Chierchia, G., Fox, D., & Spector, B. (2008). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. *Unpublished manuscript*. Harvard University and MIT, Cambridge, MA.

Condry, K. F. & Spelke, E. S. (2008). The development of language and abstract concepts: The case of natural number. *Journal of Experimental Psychology: General*, 137, 22-38.

Dehaene, S. (1997). *The number sense: How the mind creates mathematics*. Oxford, England: Oxford University Press.

Dempster, F. N. (1981). Memory span: Sources of individual and developmental differences. *Psychological Bulletin*, 89, 63-100.

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.

Foppollo, F., Guasti, M., Chierchia, G. (under review). Scalar Implicatures in Child Language: failures, strategies and lexical factors.

Fox, D. & Hackl, M. (2004). The Universal Density of Measurement. *Unpublished manuscript*. MIT, Cambridge, MA.

Frege, G. (1892). On sense and reference (trans. P. Geach and M. Black). In Geach, P. and M. Black (eds), *Translations from the Philosophical Writings of Gottlob Frege,* pp. 56-78. Oxford: Blackwell.

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

Gelman, R., & Gallistel, C. R. (1978). *The child's understanding of number*. Cambridge, MA: Harvard University Press.

Gershkoff-Stowe, L. (2002). Object naming, vocabulary growth, and the development of word retrieval abilities. *Journal of Memory and Language*, 46, 665-687.

Geurts, B. (2006). Take 'five': The meaning and use of a number word. In: Svetlana Vogeleerand Liliane Tasmowski (eds.), *Non-definiteness and Plurality*. Benjamins, Amsterdam/Philadelphia. pp 311-329.

Gilmore, C. & Spelke, E. (2008). Children's understanding of the relationship between addition and subtraction. *Cognition*, 107, 932-945.

Gordon, P. (2004). Numerical cognition without words: Evidence from Amazonia. *Science*, 306, 499-503.

Grice, H. P. (1957). Meaning. *Philosophical Review*, 66, 377-88.

Grice, H. P. (1975). Logic and Conversation. In P. Cole and J.L. Morgan (Eds.), *Syntax and Semantics*, Vol. 3., 41-58. New York: Academic Press.

Griffin, S., & Case, R. (1996). Evaluating the breadth and depth of training effects, when central conceptual structures are taught. *Monographs of the Society for Research in Child Development*, 61, 83-102.

Gualmini, A., Crain, S., Meroni, L., Chierchia, G., & Guasti, M. T. (2001). At the semantics/pragmatics interface in child language. *Proceedings of Semantics and Linguistic Theory XI*. Ithaca, NY: CLC Publications, Department of Linguistics, Cornell University.

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

Horn, L. (1972). *On the semantic properties of the logical operators in English*. Doctoral dissertation, UCLA, Los Angeles, CA. Distributed by IULC, Indiana University, Bloomington, IN.

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

Horn, L. (1992). The said and the unsaid. *Ohio State University Working Papers in Linguistics* (SALT II Proceedings) 40: 163-192.

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

Huang, Y. & Snedeker, J. (in press). Semantic meaning and pragmatic interpretation in five-year-olds: Evidence from real time spoken language comprehension. To appear in *Developmental Psychology*.

Ionin, T. & Matushansky, O. (2006). The composition of complex cardinal. *Journal of Semantics*, 23, 315-360.

Kadmon, N. (2001). *Formal Pragmatics*. Blackwell. Oxford.

Kail, R.V. (1991). Development of processing speed in childhood and adolescence. In H. W. Reese (Ed.), *Advances in child development and behavior* (Vol. 25). New York: Academic Press.

Kail, R.V. & Salthouse, T.A. (1994). Processing speed as a mental capacity. *Acta Psychologica*, 86, 199-225.

Katsos, N. & Bishop, D. (2008). A developmental investigation of the effects of scale type and speech-act on the generation of scalar implicatures. *A paper presented at the 11th Congress of the International Association for the Study of Child Language*. Edinburgh, Scotland.

Koenig, J. (1991). Scalar predicates and negation: punctual semantics and interval interpretations, *Chicago Linguistic Society 27, Part 2: Parasession on negation*, 140-155.

Kratzer, A. (1998) Scope or Pseudo-scope: Are there wide scope indefinites? In S. Rothstein (ed.) *Events and Grammar*. Kluwer, Dordrecht. 163-196.

LeCorre, M. & Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition*, 105, 395-438.

Levinson, S. (1983). Pragmatics. Cambridge: Cambridge University Press.

Levinson, S. (2000). Presumptive meanings. Cambridge, MA: MIT Press.

Li, P., Barner, D., & Huang, B. (2008). Classifiers as count syntax: Individuation and measurement in the acquisition of Mandarin Chinese. *Language Learning and Development*, 4, 1-42.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., & Xu, F. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, 57 (4, Serial No. 228).

Markman. M. (1989). *Categorization and naming in children: Problems of induction.* Cambridge, MA: MIT Press.

Musolino, J. (2004). The semantics and acquisition of number words: integrating linguistic and developmental perspectives. *Cognition*, 93, 1-41.

Noveck, I. A. (2001). When children are more logical than adults: experimental investigation of scalar implicatures. *Cognition*, 78, 165-188.

Noveck, I. A., Chierchia, G., Chevaux, F., Guelminger, R., & Sylvestre, E. (2002). Linguistic-pragmatic factors in interpreting disjunctions. *Thinking and Reasoning*, 8, 297-326.

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

Panizza, D., Chierchia, G., & Clifton, C. (in press). On the role of entailment patterns and scalar implicatures in the processing of numerals. To appear in *Journal of Memory and Language*.

Panizza, D., Chierchia, G., Huang, Y., & Snedeker, J. (2009). The relevance of polarity for the online interpretation of scalar terms. Paper presented at *Semantics and Linguistic Theory (SALT) 19*, Columbus, OH, April 2009.

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

Papafragou, A. & Tantalou, N. (2004). Children's computation of implicatures. *Language Acquisition*, 12, 71-82.

Paris, S. (1973). Comprehension of language connectives and propositional logical relationships. *Journal of Experimental Child Psychology*, 16, 278–291.

Pica. P. Lemer, C., Izard, V., & Dehaene, P. (2004). Exact and approximate arithmetic in an Amazonian indigene group. *Science*, 306, 441-443.

Pouscoulous, N., Noveck, I. A., Politzer, G., & Bastide, A. (2007). A developmental investigation of processing costs in implicature production. *Language Acquisition*, 14, 347-375.

Reinhart, T. (1999). The processing cost of reference-set computation: Guess patterns in acquisition. *OTS Working Papers* (Uil-Ots 99001-CL/TL).

Rips, L. J. (1975). Quantification and semantic memory. *Cognitive Psychology, 7*, 307-340.

Saddock, J. (1984). Whither radical pragmatics? In D. Schiffrin (ed.), *Meaning, form and use in context*. Washington: Georgetown University Press, 139-149.

Sarnecka, B. W., Kamenskaya, V. G., Yamana, Y., Ogura, T., & Yudovina, J. B. (2007). From grammatical number to exact numbers: Early meanings of *one*, *two*, and *three* in English, Russian, and Japanese. *Cognitive Psychology*, 55, 136-168.

Sarnecka, B.W. & Lee, M. D. (2009). Levels of number knowledge during early childhood. *Journal of Experimental Child Psychology*, 103, 325-337.

Scharten, R. (1997). *Exhaustive interpretation: A discourse-semantic account*. Doctoral dissertation, Katholieke Universiteit, Nijmegen.

Schneider, W., & Bjorklund, D. F. (1998). Memory. In D. Kuhn & R. S. Siegler (Eds.), Cognitive, language, and perceptual development, Vol. 2 (pp. 467-521). In W. Damon (General Editor), *Handbook of child psychology (5th Ed.).* New York: Wiley.

Schwarzschild, R. (2002). Singleton indefinites. *Journal of Semantics,* 19, 289-314.

Siegler, R.S. (2007). Cognitive variability. *Developmental Science*, 10, pp. 104-109.

Siegler, R.S., & Shrager, J. (1984). Strategy choices in addition and subtraction: how do children know what to do? In C. Sophian (Ed.), *The origins of cognitive skills* (pp. 229–293). Hillsdale, NJ: Erlbaum.

Smith, C. L. (1980). Quantifiers and question answering in young children. *Journal of Experimental Child Psychology*, 30, 191-205.

Snedeker, J. & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology*, 49, 238-299.

Strawson, P. F. (1950). On referring. *Mind,* 59, 320-44.

Swingley, D. & Fernald, A. (2002). Recognition of words referring to present and absent objects by 24-month-olds. *Journal of Memory and Language*, 46, 39-56.

Syrett, K., Kennedy, C., & Lidz, J. (in press). Meaning and context in children's understanding of gradable adjectives. To appear in *Journal of Semantics.*

Trueswell, J.C., Sekerina, I., Hill, N.M. & Logrip, M.L. (1999). The kindergarten path effect: studying on-line sentence processing in young children. *Cognition*, 73, 89-134.

van Rooy, R. & Schulz, K. (2006). Pragmatic meaning and non-monotonic reasoning: The case of exhaustive interpretation. *Linguistics and Philosophy*, 29, 205–250.

Winter, Y (2001). *Flexibility principles in boolean semantics*. Cambridge, Mass: MIT Press.

Wynn, K. (1990). Children's understanding of counting. *Cognition*, 36, 155-193.

Wynn, K. (1992). Children's acquisition of the number words and the counting system. *Cognitive Psychology*, 24, 220-251.

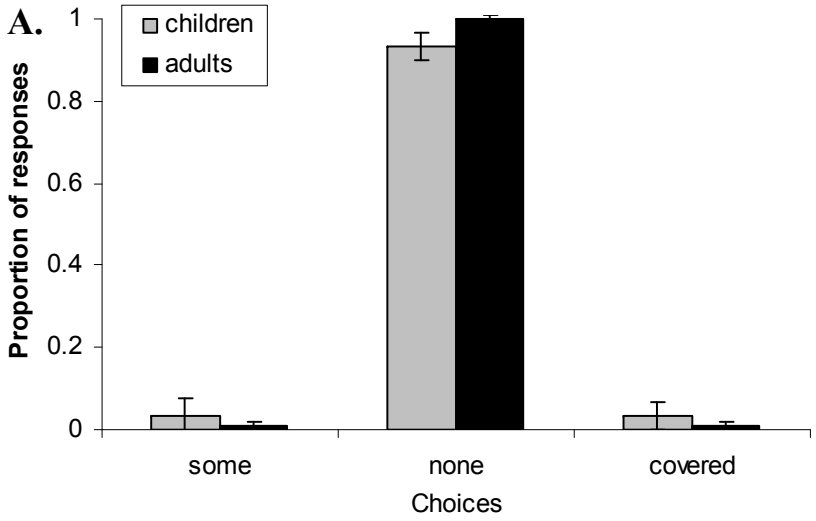
Zur, O. & Gelman, R. (2004). Young children can add and subtract by predicting and checking. *Early Childhood Research Quarterly,* 19, 121–137.

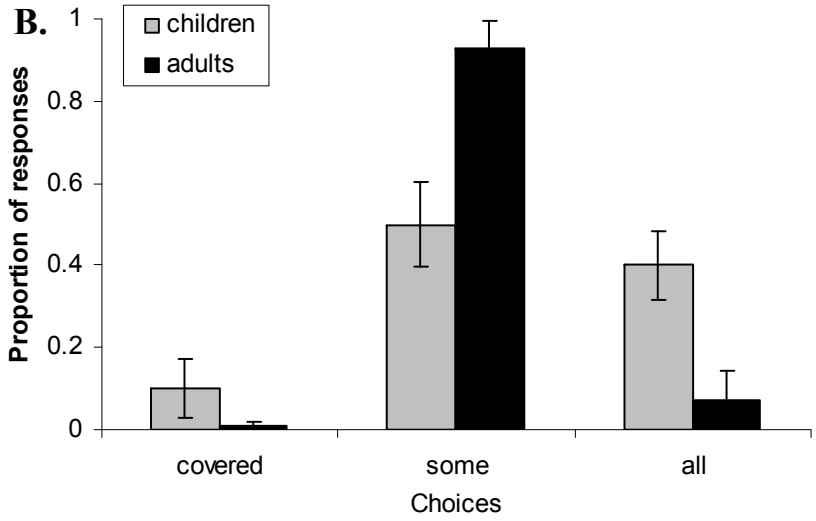
Table 1: Children's performance on the highest number that they are credited with knowing in the Give-N Task (N) and the next highest number (N+1) based on the data from Sarnecka and colleagues' study of children learning three languages (2007).

		Knowledge of N		**Knowledge of N+1**	
		Hit Rate	**False Alarm Rate**	**Hit Rate**	**False Alarm Rate**
		(N for N)	*(N for more)*	*(N+1 for N+1)*	*(N+1 for more)*
0-Knower	English (4)	-	-	0.75	0.57
	Japanese (22)	-	-	0.70	0.58
	Russian (1)	-	-	1.00	0.67
	Average	-	-	0.72	0.57
1-Knower	English (26)	0.95	0.01	0.73	0.63
	Japanese (16)	0.96	0.01	0.63	0.52
	Russian (20)	0.98	0.00	0.67	0.58
	Average	0.96	0.01	0.68	0.59
2-Knower	English (13)	0.95	0.03	0.49	0.33
	Japanese (5)	0.80	0.04	0.53	0.30
	Russian (11)	0.94	0.05	0.46	0.23
	Average	0.92	0.04	0.48	0.29
Average	English	0.95	0.02	0.66	0.54
	Japanese	0.92	0.02	0.65	0.53
	Russian	0.97	0.02	0.60	0.46
Grand Average	1 & 2-knowers	0.95	0.02	0.62	0.49
	All	-	-	0.64	0.51

Figure 1: In the scalar task, participants were presented with boxes where Big Bird (on the left) and Cookie Monster (on the right) possessed cookies and asked to "Give me the box where Cookie Monster has some of the cookies" in (A) NONE vs. SOME, (B) ALL vs. SOME, and (C) NONE vs. ALL trials.

Figure 2: When asked for *some* (A) in the NONE vs. SOME trials, participants selected the box with SOME of the cookies. (B) In the ALL vs. SOME trials, adults selected the box with SOME of the cookies while children were equally selected either the box with SOME or ALL of the cookies. (C) Finally, in the NONE vs. ALL trials, participants selected the box with ALL of the cookies.





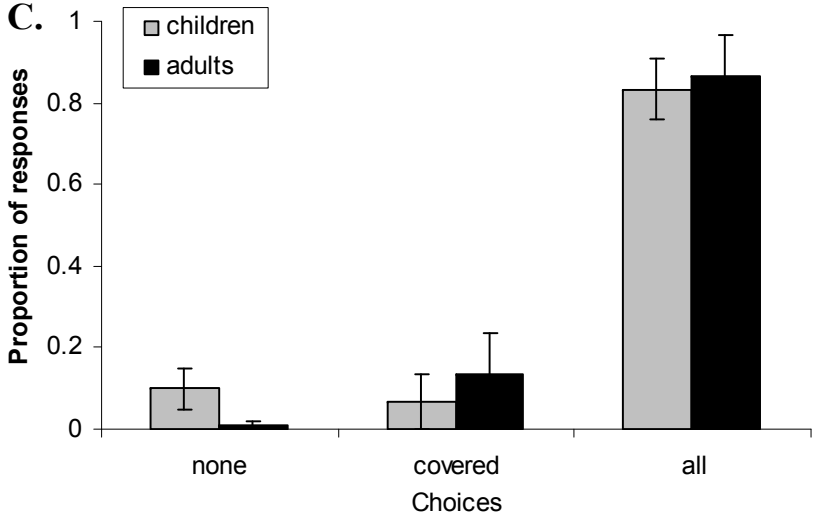


Figure 3: In the number task, participants were asked to *"Give me the box with two fish"* for the (A) EXACT VS. LESS, (B) EXACT VS. MORE, and (C) the LESS VS. MORE trials.

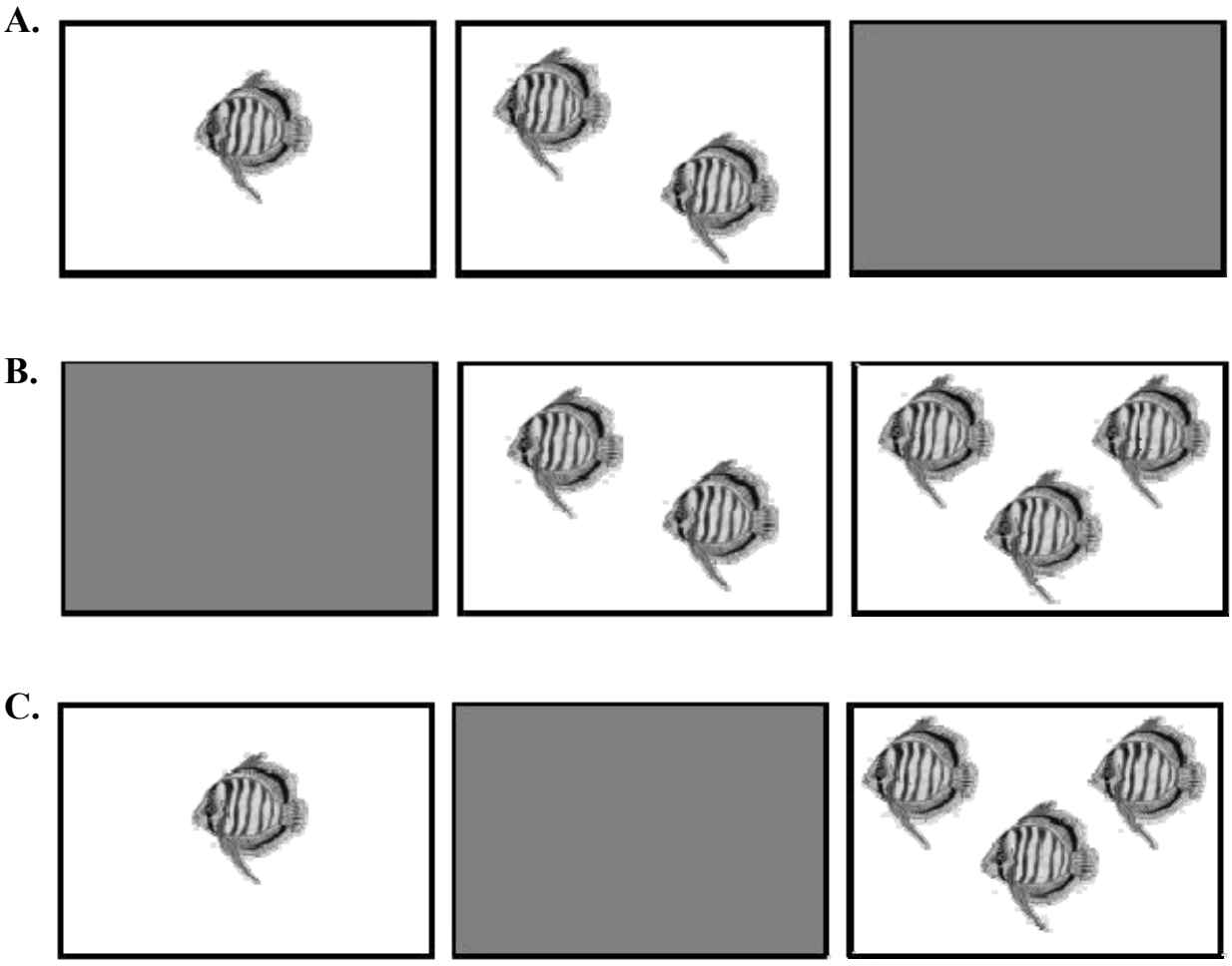
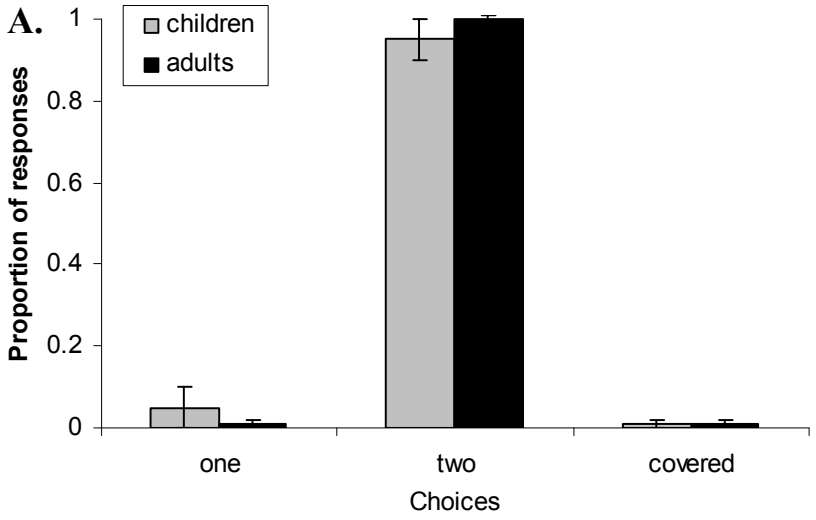
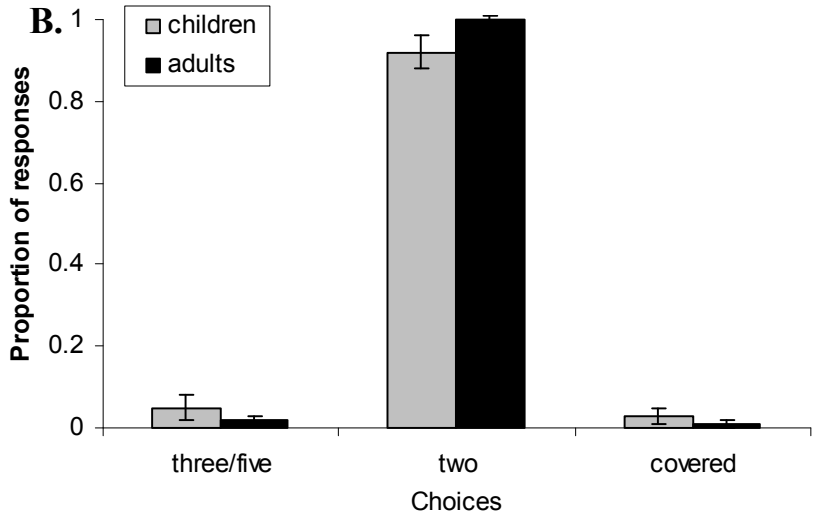


Figure 4: When asked for *two*, participants overwhelmingly selected the exact match (A) in the EXACT vs. LESS trials and (B) in the EXACT vs. MORE trials. (C) However, in the critical LESS vs. MORE trials, participants selected the covered box.





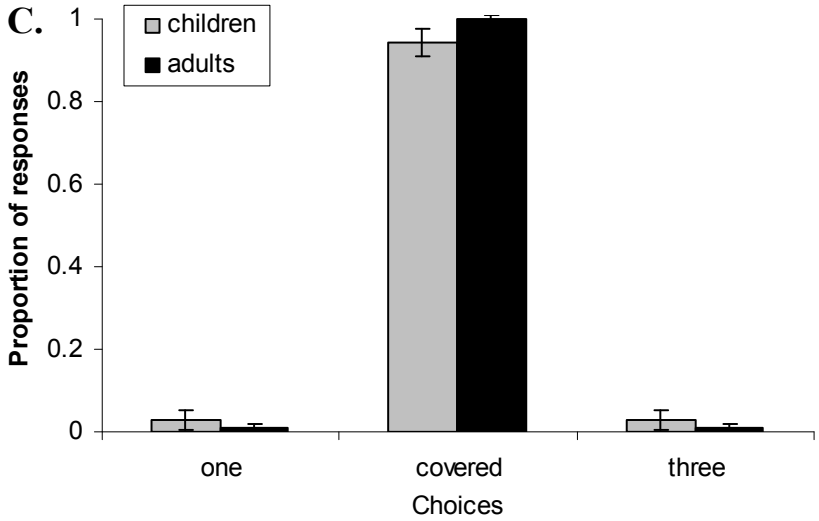
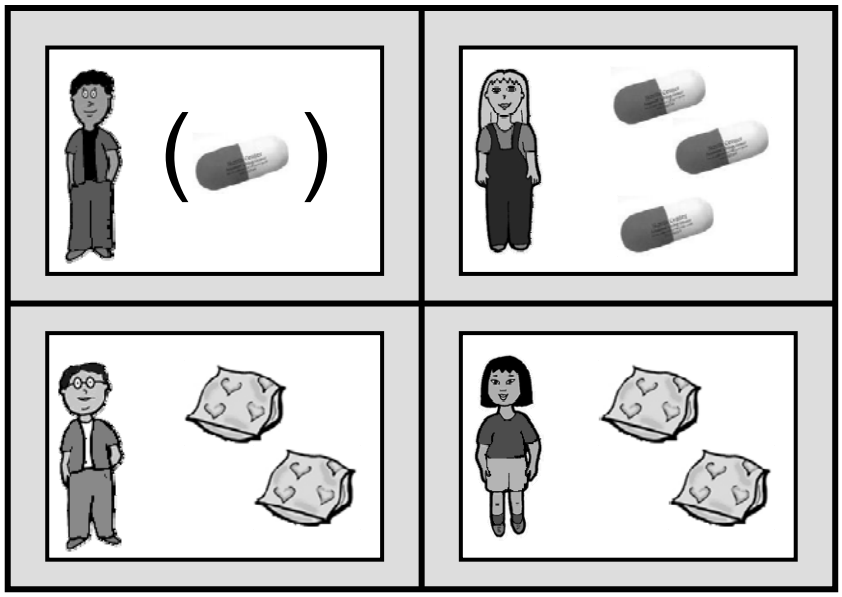


Figure 5: In the eye-tracking task, examples of displays for (A) *all* and *three* trials and (B) *some* and *two* trials. Participants here were instructed to *"Point to the girl that has ____ of the pills."* An additional object (in parentheses) was added to the character of opposite gender makes the utterance felicitous for number trials. The girl with pills was the Target while the girl with pillows was the Distractor.

A.



B.

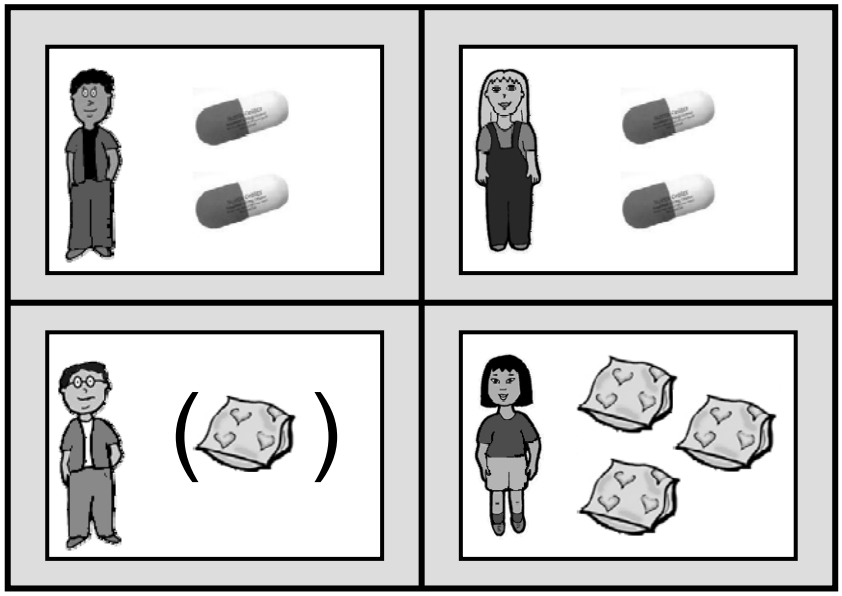


Figure 6: In the eye-tracking task, the time-course of adults' looks to Target for the four trial types.

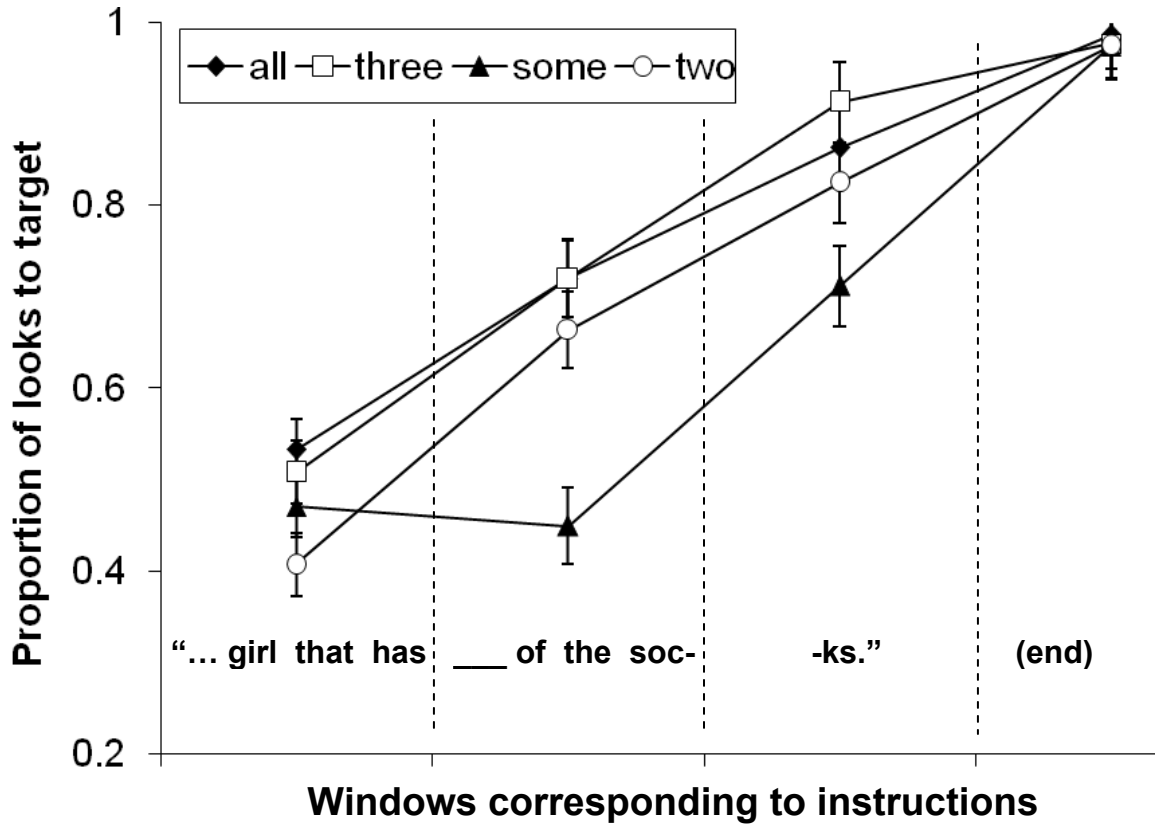


Figure 7: Trials were separated based on fixations prior to the onset of the quantifier. (A) From the 600 ms to 800 ms time window, adults in the Non-Target initial trials were more likely to switch their looks to the Target following *two*, *three*, and *all* than they were for *some*. (B) At the 200 ms time window, adults in the Target initial trials were less likely to abandon their looks to the Target following *two*, *three*, and *all* than they were for *some*.

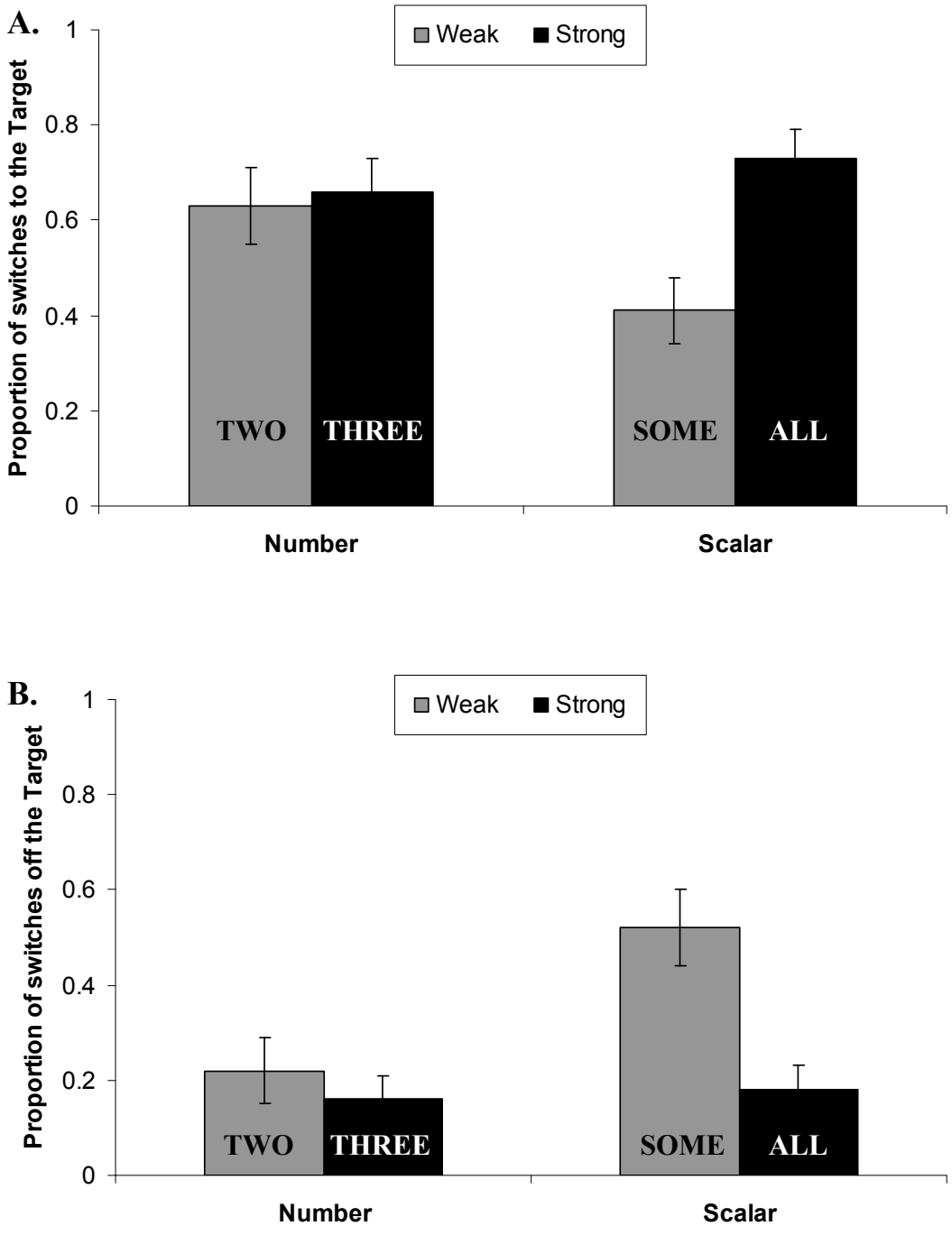


Figure 8: In the eye-tracking task, the time-course of children's looks to Target for the four trial types.

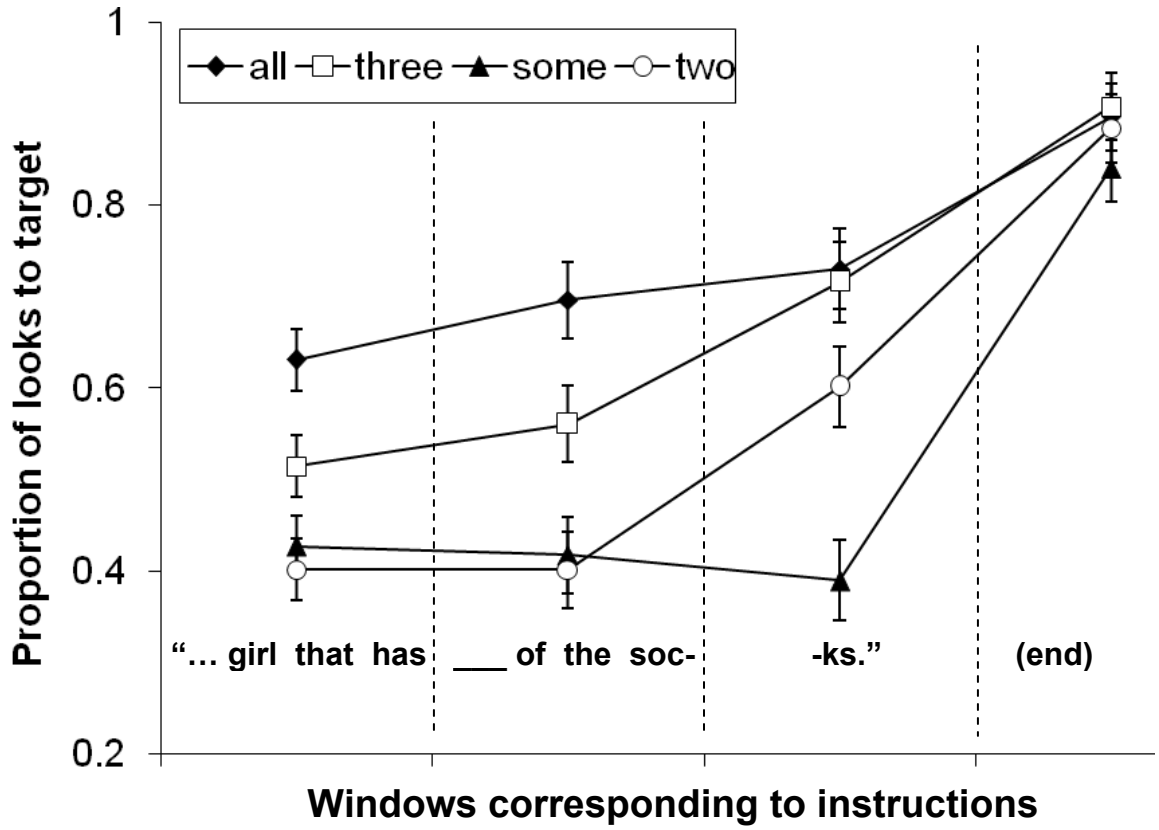


Figure 9: Trials were separated based on fixations prior to the onset of the quantifier. (A) At the 800 ms time window, children in the Non-Target initial trials were more likely to switch their looks to the Target following *two*, *three*, and *all* than they were for *some*. (B) From the 500 ms to 700 ms time window, children in the Target initial trials were less likely to abandon their looks to the Target following *two*, *three*, and *all* than they were for *some*.