2-1-2001

Scalar Implicatures: Experiments at the Semantics-Pragmatics Interface

Anna Papafragou
University of Pennsylvania, anna4@linc.cis.upenn.edu

Julien Musolino
Indiana University, Bloomington, musolino@indiana.edu

University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-01-15.

Scalar Implicatures: Experiments at the Semantics-Pragmatics Interface

Abstract
In this article we present two sets of experiments designed to investigate the acquisition of scalar implicatures. Scalar implicatures arise in examples like *Some professors are famous* where the speaker's use of *some* typically indicates that s/he had reasons not to use a more informative term, e.g. *all*. *Some professors are famous* therefore gives rise to the implicature that not all professors are famous. Recent studies on the development of pragmatics suggest that preschool children are often insensitive to such implicatures when they interpret scalar terms (Noveck 2001 for terms like *might* and *some*; Chierchia, Crain, Guasti, Gualmini and Meroni 2001 for *or*). This conclusion raises two important questions: a) are all scalar terms treated in the same way by young children?, and b) does the child's difficulty reflect a genuine inability to derive scalar implicatures or is it due to demands imposed by the experimental task on an otherwise pragmatically savvy child? Experiment 1 addresses the first question by testing a group of 30 5-year-olds and 30 adults (all native speakers of Greek) on three different scales, *meriki/oli (some/all), dio/tris (two/three)* and *arxizo/teliono (start/finish)*. In each case, subjects were presented with contexts which satisfy the truth conditions of the stronger (i.e. more informative) terms on each scale (i.e. *all, three* and *finish*) but were described using the weaker terms of the scales (i.e. *some, two, start*). We found that while adults overwhelmingly rejected these infelicitous descriptions, children almost never did so. Children also differed from adults in that their rejection rate on the numerical scale was reliably higher than on the two other scales. In order to address question (b), we trained a group of 30 5-year-olds to detect infelicitous statements. We then presented them with modified versions of the stories of Experiment 1, which now more readily invited scalar inferences. These manipulations gave rise to significantly higher rejection rates than those observed in Experiment 1. Overall, these findings indicate that children do not treat all scalar terms alike and, more importantly, that children's ability to derive scalar implicatures is affected by their awareness of the goal of the task. Developmental and methodological implications as well as theoretical implications for the semantics of numeral terms are discussed.

Keywords
language acquisition, pragmatics, sentence processing, scalar implicatures, quantifiers, numerals, Greek

Comments
University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-01-15.

Scalar Implicatures: Experiments at the Semantics-Pragmatics Interface

ANNA PAPAFRAGOU[a] & JULIEN MUSOLINO[b]

[a]*Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia PA 19104*

[b]*Department of Speech and Hearing Sciences, Indiana University, Bloomington IN 47405*

Abstract

In this article we present two sets of experiments designed to investigate the acquisition of scalar implicatures. Scalar implicatures arise in examples like *Some professors are famous* where the speaker's use of *some* typically indicates that s/he had reasons not to use a more informative term, e.g. *all*. *Some professors are famous* therefore gives rise to the implicature that not all professors are famous. Recent studies on the development of pragmatics suggest that preschool children are often insensitive to such implicatures when they interpret scalar terms (Noveck 2001 for terms like *might* and *some*; Chierchia, Crain, Guasti, Gualmini and Meroni 2001 for *or*). This conclusion raises two important questions: a) are all scalar terms treated in the same way by young children?, and b) does the child's difficulty reflect a genuine inability to derive scalar implicatures or is it due to demands imposed by the experimental task on an otherwise pragmatically savvy child? Experiment 1 addresses the first question by testing a group of 30 5-year-olds and 30 adults (all native speakers of Greek) on three different scales, *meriki/oli (some/all)*, *dio/tris* (*two/three*) and *arxizo/teliono* (*start/finish*). In each case, subjects were presented with contexts which satisfy the truth conditions of the stronger (i.e. more informative) terms on each scale (i.e. *all*, *three* and *finish*) but were described using the weaker terms of the scales (i.e. *some*, *two*, *start*). We found that while adults overwhelmingly rejected these infelicitous descriptions, children almost never did so. Children also differed from adults in that their rejection rate on the numerical scale was reliably higher than on the two other scales. In order to address question (b), we trained a group of 30 5-year-olds to detect infelicitous statements. We then presented them with modified versions of the stories of Experiment 1, which now more readily invited scalar inferences. These manipulations gave rise to significantly higher rejection rates than those observed in Experiment 1. Overall, these findings indicate that children do not treat all scalar terms alike and, more importantly, that children's ability to derive scalar implicatures is affected by their awareness of the goal of the task. Developmental and methodological implications as well as theoretical implications for the semantics of numeral terms are discussed.

Key words: Language acquisition, pragmatics, sentence processing, scalar implicatures, quantifiers, numerals, Greek.

1. Introduction

Humans are extremely sophisticated communicators. Perhaps the most striking demonstration of the subtle and complex nature of human communication is the fact that speakers often intend to convey far more than the words they utter encode and hearers manage to go beyond what the speaker has uttered and retrieve the intended interpretation of the utterance. In other words, although human language consists of essentially arbitrary (code-like) pairings between sound and meaning, verbal communication involves much more than a simple encoding-decoding procedure: it crucially involves inference.

This paper focuses on one of the best-known types of pragmatic inference. Consider the following example:

(1) A: Some top models are rich.

Most people would agree that, in producing the utterance in (1), the speaker conveys the assumption in (2):

(2) Not all top models are rich.

Notice that (2) is not encoded by the speaker's utterance, nor is it part of what the speaker has *said*: according to standard semantic accounts, *some* means *some and possibly all*. Rather, (2) is an assumption inferentially derived on the basis of what the speaker said. The intuition behind this interpretation has long been familiar to philosophers (see Horn 1992):

> If I say to any one, "I saw some of your children today", he might be justified in inferring that I did not see them all, not because the words mean it, but because, if I had seen them all, it is most likely that I should have said so: even though this cannot be presumed unless it is presupposed that I must have known whether the children I saw were all or not. (Mill 1867: 501)

> Whenever we think of the class as a whole, we should employ the term All; and therefore when we employ the term Some, it is implied that we are not thinking of the whole, but of a part as distinguished from the whole - that is, of a part only. (Monck 1881: 156)

Even though the idea behind the reasoning was clear, the first systematic attempt to explain how the inference in (2) is derived belongs to Paul Grice. As is well known, Grice offered a comprehensive framework of the mechanics of inferential communication (Grice 1989). He suggested that communication is essentially a co-operative enterprise governed by certain rational expectations ('Maxims') about how a conversational exchange should be conducted. According to Grice's maxims, interlocutors are normally expected to offer contributions which are truthful, informative, relevant to the goals of the conversation and appropriately phrased. These expectations about rational conversational conduct constrain the range of inferences which hearers are entitled to entertain when interpreting utterances. Furthermore, these expectations can be violated (or exploited) to create a variety of effects. According to Grice's scheme, in producing (1), the speaker has violated the maxim of informativeness (or Quantity):

Quantity maxim

i. Make your contribution as informative as is required.

ii. Do not make your contribution more informative than is required.

Specifically, the speaker has violated the submaxim (i) since s/he has chosen a relatively weak term from among a range of items ordered in terms of informational strength (<*all,…, some*>). Assuming that the speaker is trying to be co-operative and will say as much as she truthfully can that is relevant to the exchange, the fact that s/he chose the weaker term (*some*) gives the listener reason to think that she is not in a position to offer an informationally stronger statement (*All top models are rich*). This leads to the inference that the stronger statement is false, i.e. to (2). The assumption in (2) is a (conversational) *implicature* - more specifically, a Quantity or scalar implicature (so called because of the informational scale <*some, all*>; Horn 1972).[1]

[1] Scalar implicatures can be cancelled; cf. *Some top models are rich - in fact all of them have tons of money.*

Examples such as (1) have been studied extensively in post-Gricean philosophical and linguistic work on verbal communication. Much of this work has tried to tease apart in a principled way the contribution of semantics (what is linguistically encoded) vs. pragmatics (what is inferred on the basis of the linguistically encoded meaning) in utterance interpretation. Within this larger project, scalar implicatures have been considered by many authors as the paradigm case of pragmatic inference and an important testing ground for competing pragmatic theories. As a result, several accounts of scalar inferences have been developed beyond Grice's original proposal (Horn 1972, Harnish 1975, Gazdar 1979, Hirschberg 1985, Carston 1995, Sperber and Wilson 1995, Levinson 2000).

Given their importance for pragmatic theory, scalar inferences raise a host of interesting psychological issues. First, are scalar implicatures psychologically real? Do competent hearers take them into account when processing utterances? And is their on-line derivation properly captured by the mechanisms described by pragmatic theories? Few studies have attempted to address the psychology of implicature (see e.g. Clark 1992, Gibbs 1994, Cacciari and Glucksberg 1994) and only recently have scalar terms attracted experimental attention (Noveck 2001). These studies have offered support for the intuition that implicatures play a central part in mature verbal communication. Some of the most interesting evidence comes from reasoning studies (Begg and Harris 1982, Newstead and Griggs 1983), which show how the presence of scalar implicatures disrupts the expected performance of subjects in standard logical tasks. However, the precise mechanisms underlying the calculation of these implicatures have not been systematically explored.

The pragmatics of scalar terms also raise serious developmental questions: How is the ability to compute scalar inferences acquired? More generally, how do young learners become capable of deriving implicated meanings? Recent experimental investigations into children's interpretation of scalar terms have concluded that preschool children are often insensitive to scalar implicatures in tasks involving language comprehension (Noveck 2001, Chierchia, Crain, Guasti, Gualmini and Meroni 2001). Children in these studies, although otherwise linguistically competent, were shown to attend only to the logical meaning of scalar elements such as *some* and *or* rather than the inferred meaning (i.e. *some but not all; A or B but not both*). Such behavior has led Noveck (2001) to conclude that "younger, albeit competent reasoners, initially treat a relatively weak term logically before becoming aware of

its pragmatic potential", and that, in this respect, "children are more logical than adults" (Noveck 2001: 165).

These findings are reminiscent of a long line of results showing that young children fail in various pragmatic-inferential tasks, from the resolution of structural ambiguity (Trueswell, Sekerina, Hill and Logrip 1999) to the comprehension of metaphor and irony (Shatz 1980; for further examples, see Markman and Seibert 1976, Gelman and Greeno 1989). If it is true that young children have difficulty with scalar inferences, it is worth asking how and when children become capable of computing them. Alternatively, one may try to produce evidence that children are aware of the pragmatic potential of scalar expressions; if successful, this approach should also lead to some initial insights into the conditions under which children exhibit this kind of pragmatic sophistication.

In this paper, we report results from two sets of experiments designed to investigate the computation of scalar implicatures by both children and adults. Our studies have two main goals. For adults, we aim at verifying that scalar implicatures are produced regularly and in the contexts described by pragmatic theory. To the extent that this is the case, we then ask whether children are capable of producing scalar implicatures and under what conditions. Anticipating our findings, we show that children's success with scalar inferences depends crucially on the semantics of individual scalar terms and is subject to context effects: whenever context and semantics conspire in appropriate ways, children are capable of computing scalar implicatures in almost adult-like fashion. We also conclude that child data are informative about the semantic representation of scaleable terms (e.g. number terms such as *two* and *three*) and can therefore be used as evidence in theoretical debates over the semantics-pragmatics distinction.

One of the properties of pragmatic inference is its universality: Since implicatures are motivated, not arbitrary, we expect them to arise cross-linguistically in much the same way. Our studies tested the comprehension of scalar terms in Modern Greek, fully expecting that our adult subjects would compute scalar inferences in ways no different from English speakers. This expectation was borne out in our results.

To preface the experimental part of the paper, we present some pragmatic background to scalar implicature and we introduce a variety of expressions and contexts which license such inferences (Section 2). We go on to review in some detail previous experimental studies

of scalar implicatures (Section 3). These will provide the theoretical and methodological backdrop to the experiments which form the main part of the paper (Sections 4 and 5).

2. Background on the pragmatics of scalar terms

Recall that the main idea behind scalar inference is that the hearer evaluates what the speaker has uttered against a set of ordered alternatives (a *scale*), given certain expectations about what the speaker wished to communicate. In the right circumstances, the assertion of the lower ranking alternative implicates that the speaker is not in a position to assert the higher ranking one.[2] Beyond the familiar case of quantifiers such as <*all, some*>, classic examples of scales include numerals (<…*three, two, one*>), modals (<*necessarily, possibly*>, <*must, should, may*>), connectives (<*and, or*>), adverbs (<*always, often, sometimes*>), degree adjectives (<*hot, warm*>) and verbs of ranking (<*know, believe*>, <*love, like*>) or completion (<*start, finish*>). What defines these information scales, in both logical and nonlogical vocabulary, is the presence of one-way semantic entailments (e.g. for an arbitrary simplex sentence-frame S, *S(all)* entails *S(some)*, but not vice versa):

(3) A: Do you have three extra paper clips?
B: I have two.
→ B doesn't have three extra paper clips.

(4) A: Do you think the Yankees will win?
B: It's possible.
→ It's not certain that the Yankees will win.

(5) A: What do you do on weekends?
B: I sometimes play basketball.

[2] There is, in fact, considerable debate over the precise epistemic commitment communicated by scalar implicatures: *Some came* may imply that the speaker doesn't know whether all came, or knows that not all came (Levinson 2000: 77-79). Moreover, the utterance may in other contexts imply that the speaker does not want to say whether all came (see the Postface in Sperber & Wilson 1995). We will not be concerned with these

→ B doesn't always play basketball on weekends.

(6) A: Are you done with your 'Psychology of Drama' essay?
B: I started writing it.
→ B didn't finish writing her 'Psychology of Drama' essay.

Scalar orderings are also licensed by a variety of other relations beyond entailment (Fauconnier 1975, Hirschberg 1985). For instance, several vocabulary items which enter into some form of lexical contrast may license scalar inferences: the statement that someone *tried* to do something usually implicates that they didn't *succeed*, even though *try* and *succeed* do not form an entailment scale. More generally, scalar inferences can be induced by partial contextual orderings, which may be supplied by stable world knowledge or created in a completely *ad hoc* fashion (examples from Hirschberg 1985):[3]

(7) A: Are John and Mary married?
B: They're engaged.
→ They are not married.

(8) A: Did you get Paul Newman's autograph?
B: I got Joanne Woodward's.
→ B didn't get Paul Newman's autograph.

Two properties of scalar inferences are particularly relevant for our present purposes. First, there is considerable debate as to whether all scalar orderings form a homogeneous class, even within the range of entailment scales. It has recently been argued (Carston 1990, 1998; cf. also Horn 1992) that cardinals behave differently from other scalar terms in important ways. For instance, it is possible to use numbers with an 'at most' reading (*She can have 2000 calories a day without putting on weight*), while no such reading is available with other

---

important issues in this paper. For present purposes, we assume that scalar implicatures come without propositional attitude markers: in the example above, the speaker will be taken to imply simply *Not all came*.

[3] Entailment scales can thus be viewed as the limiting case of a spectrum of different contextually determined orderings of alternates, where both the elements within these orderings and their relation are pragmatically specified.

scalar terms (e.g. *some* cannot be interpreted as 'at most some'). Similarly, number terms are regularly used with an 'exact' interpretation (e.g. in mathematical statements such as *Two plus two makes four*), a fact which is hard to explain if number terms have a (Gricean-Hornian) 'at least' semantics. Further, cardinal expressions are ordinarily used with an 'exact' interpretation in compounds: a three-sided figure cannot be a figure with at least three sides (e.g. a square), unlike what an 'at least' semantics for cardinals appears to predict. On the basis of these and similar arguments, it has been argued that numerals may not have an 'at least' semantics which is then upper-bounded by a scalar implicature; rather they might be best analyzed as underspecified among the 'at least', 'exact' and 'at most' readings. If this line of reasoning is correct, the scalar inferences associated with numerals would no longer be considered conversational *implicatures* but would come out as different ways of pragmatically enriching the underspecified semantic content of the numerals (for further discussion, see Sadock 1984, Carston 1988, Levinson 2000, and many others).

A second property of scalar inference is that it is subject to the demands of the communicative exchange. Notice that, without any further specifications, the system described so far would massively overgenerate scalar inferences. But naturally, not all weak propositions implicate the negation of a stronger one. Whether a scalar inference will be produced in a certain instance is constrained by relevance (Sperber and Wilson 1995, Carston 1995, 1998).[4] The main consideration here lies with expectations of cognitive effect: If the communicative expectations of the hearer are met in the 'weaker' statement, no inference should arise. So *Some like it hot* does not imply *Not everybody likes it hot* simply because the former satisfies the hearer's communicative expectations (i.e. it gives rise to adequate cognitive gains). However, if what the speaker has said falls short of these expectations (i.e. is 'informationally weaker'), the hearer is entitled to go beyond what was said and draw the inference that a 'stronger' proposition does not hold.

We return to these properties of scalar phenomena in the discussion and experimentation which follow. We ask, among other things, whether all scales are equal in developmental terms, paying special attention to number scales. We also look at ways in which

[4] Recently, certain commentators have proposed (expanding on Grice 1989) that scalar implicatures belong to a more general class of default conversational inferences which are present more or less regardless of context, provided that an appropriate term (e.g. *some*) is used (Levinson 2000). For relevant discussion, see Carston (1995), Hirschberg (1985).

manipulations of context (and of expectations of cognitive effects) affect the computation of implicatures within a single scale.

3. Developmental studies of scalar implicature: A review

Although it is only very recently that the development of scalar implicature has begun to be experimentally investigated, a number of earlier studies designed to investigate children's knowledge of quantification (Smith, 1980) and propositional connectives (Braine and Rumain, 1981) uncovered a rather intriguing set of findings which, in retrospect, can be seen as being directly relevant to the study of scalar implicature. For example, Carol Smith (1980) discovered that preschool children, who had mastered many of the syntactic aspects of quantifiers like *some* and *all*, nevertheless often treated *some* as being compatible with *all* when asked questions such as 'Do some elephants have trunks?'. In a similar vein, Braine and Rumain (1981) found that while adult subjects tended to favor an exclusive interpretation of the disjunction operator *or* (i.e. 'A or B but not both'), children favored the so-called 'logical' interpretation of *or* on which 'A or B' is compatible with 'A and B' (see Paris, 1973 for reports of a similar developmental pattern).

Even though earlier findings are clearly relevant, the first systematic investigation of the development of scalar implicature can be attributed to Noveck (2001). Using a modal reasoning scenario, Noveck investigated children and adults' interpretations of statements expressing *x might be y* in contexts in which the stronger statements *x must be y* were true. As Noveck points out, *x might be y* can be interpreted logically (as compatible with *must*) or pragmatically (as exclusive to *must*). Noveck's main finding is that 7 to 9-year-old children treated *x might be y* logically (i.e., as compatible with *x must be y*) much more often than adults. Noveck also reports that 8 to 10-year-old children treated French *certains* ('some') as compatible with *tous* ('all') much more often than adults.

In a related set of recent studies, Chierchia, Crain, Guasti, Gualmini and Meroni (2001) and Gualmini, Crain, Meroni, Chierchia and Guasti (2001) investigated preschooler's interpretation of the disjunction operator, *or*, in contexts which give rise to the implicature of exclusivity (e.g., *Every boy chose a skateboard or a bike*) and in contexts in which the implicature does not arise (e.g., *Every dwarf who chose a banana or a strawberry received a jewel*). As in previous

studies, these authors found that while adults were sensitive to the implicature in contexts in which it was predicted to arise, some of the children they tested showed virtually no sensitivity to it. In order to determine whether children would prefer a statement containing *and* (e.g., *Every farmer cleaned a horse and a rabbit*) over a statement containing *or* (e.g., *Every farmer cleaned a horse or a rabbit*) in a situation in which all the farmers had cleaned both a horse and a rabbit, Chierchia et al. presented children with both statements (produced by two puppets) and asked them to reward the puppet who 'said it better'. What was found here is that children overwhelmingly chose to reward the puppet who had produced the statement containing the conjunction *and*. Commenting on the results presented by Chierchia et al., Gualmini et al. (2001) conclude that "The explanation offered by Chierchia et al. for this set of findings is that children have knowledge of the relative information strength of sentences with *or* versus ones with *and*, and they also use information strength as the basis of their preference for sentences with *and* … Children do seem unable, however, to construct the relevant alternatives on-line, such that they fail to compute implicatures if the alternatives are not explicitly presented to them." (p. 11)

Finally, in another recent study, Musolino and Lidz (2001) report that when given a Truth Value Judgment Task, children almost always treat sentences containing *not every* (e.g., *The strong guy didn't put every elephant on the table*) as compatible with situations in which the strong guy didn't put ANY of the elephants on the table. By contrasts, adult speakers almost never do so.

The developmental picture emerging from the studies described above suggests that preschoolers – who are otherwise linguistically competent – generally perform poorly on tasks in which they are asked to assess underinformative statements. As Noveck (2001) observes, children initially appear to be 'more logical than adults' in the sense that "These results … reveal a consistent ordering in which representations of weak scalar terms tend to be treated logically by young competent participants and more pragmatically by older ones." (p. 165). This conclusion raises two important questions. The first concerns the scope of the phenomenon: do children experience equal difficulty with all scalar terms or is this difficulty restricted to the kinds of scales studied so far? In other words, do preschoolers treat all scalar terms alike? Second, one may wonder why linguistically competent children so often fail to derive scalar implicatures. One possibility is that this failure reflects a genuine inability to engage in the computations required to derive scalar implicatures. Another possibility is

that this failure may be due to the demands imposed by the experimental task on an otherwise pragmatically savvy child. To quote Noveck again, "The tasks described here, which are typical of those found in the developmental literature, demand no small amount of work as they require children to compare an utterance to real world knowledge. This might well mask an ability to perform pragmatic inferencing at younger ages." (p. 184). If so, it may be worth asking under what experimental circumstances children's ability to derive scalar implicatures may improve.

The studies reported in the present article address the two fundamental questions raised above. First, in Experiment 1, we investigated the scope of children's difficulty with scalar terms by testing their performance (as well as that of adult subjects) on three kinds of scale: a) a scale involving quantificational expressions, i.e. *some*/*all*; b) a scale involving number terms, i.e. *two*/*three,* and c) a scale involving inchoative/completion predicates, i.e. *start*/*finish*. Second, in order to determine whether children's ability to derive scalar implicatures can improve, we tested another group of children on the three scales described above under different experimental conditions (Experiment 2). Our general goal was to produce a map of pragmatic development which is not only broader than the one provided by previous studies but also one which is more detailed in the sense that it directly aims at isolating the factors susceptible to affect the child's performance on tasks designed to assess pragmatic abilities.

Our choice of scales was motivated in part by theoretical considerations. We wanted to have a range of different expressions spanning logical and nonlogical vocabulary, including numeral scales. Recall from the previous section that some of the semantics/pragmatics literature now proposes to regard numerals as different from other scalar terms (see section 2). It would therefore be of considerable theoretical interest to determine whether such differences are reflected in development. In addition to these theoretical considerations, our choice of scales was also motivated by practical questions such as the need to use scalar terms that can be easily tested given our experimental paradigm, or the desire to avoid scales that rely on idiosyncratic information. For instance, although it would be interesting to include purely pragmatic scales in a comprehensive test for scalar implicatures - see examples (7) and (8) above - such inferences are fragile and depend heavily on encyclopedic information. Unless it is independently confirmed that children of this age have the relevant background assumptions, it may not appropriate to use those scales to test for scalar reasoning in young children.

4. Experiment 1

In this first experiment, we tested children and adult speakers of Greek on their interpretation of three kinds of scalar terms: *meriki* (*some*), *dio* (*two*) and *arxizo* (*start*), used in sentences like the ones below:[5]

(9) Merika apo ta aloga pidiksan pano apo to fraxti.
some of the horses jumped over of the fence
'Some of the horses jumped over the fence.'

(10) Dio apo ta aloga pidiksan pano apo to fraxti.
two of the horses jumped over of the fence
'Two of the horses jumped over the fence.'

(11) To koritsi arxise na ftiaxni to pazl.
the girl started making the puzzle
'The girl started making the puzzle.'

In each case however, these sentences were used to describe situations which satisfied the truth conditions of sentences containing stronger terms on the respective scales, i.e., *all*, *three* and *finish*:

(12) *All* of the horses jumped over the fence.
(13) *Three* of the horses jumped over the fence.
(14) The girl *finished* making the puzzle.

So, for example, in the story corresponding to (9), three horses tried to jump over a fence (Figure 1) and all three of them managed to do so (Figure 2). Describing this situation by using a sentence like (9) is therefore pragmatically infelicitous, albeit semantically true, because the use of *merika* ('some') in (9) implies that not all of the horses jumped over the

[5] For ease of exposition, we henceforth use the English glosses of the Greek examples and terms throughout.

fence. To the extent that subjects are sensitive to this implicature, they should judge (9) to be a 'bad' description of a situation in which all of the horses jumped over the fence.[6]

INSERT FIGURES 1 & 2

Following this design strategy, the example in (10) was used as a description of a situation in which three horses (not two) jumped over a fence, and the example in (11) was used to describe a situation in which the girl in question not only started making the puzzle but in fact finished making it. Subjects' judgments of statements like (9-11) used to describe situations that satisfy the truth conditions of statements like (12-14) can therefore be taken as a measure of their sensitivity to scalar implicatures. Specifically, we expect pragmatically savvy subjects to conclude that statements like (9-11) are not a good way of describing situations in which (12-14) are true.

Method

Subjects

The participants in this experiment were a group of 30 Greek-speaking 5-year-olds (10 boys and 20 girls) between the ages of 4;11 and 5;11 (mean 5;3) and a group of 30 adult native speakers of Greek. We chose to look at preschoolers because previous studies on children's ability to derive scalar inferences have typically focused on this age group (see studies described in Section 3 above). The children who participated in this study were recruited from daycares in the Athens area. The adults were all undergraduate students at the University of Athens.

[6] Notice that our examples make use of the partitive genitive construction. This involves quantification over specific domains and encourages a contrast between, e.g., *some of the horses* and *all of the horses* (compare *some horses* and *all horses*). We expected that the presence of the partitive genitive would help our younger subjects compute the scalar implicature.

Procedure and Materials

We asked children (and adults) to judge sentences containing the scalar terms in (9-11) using a slightly modified version of the Truth Value Judgment Task Methodology (Crain and Thornton 1998). The TVJT typically involves two experimenters. The first experimenter acts out short stories in front of the subjects using small toys and props. The second experimenter plays the role of a puppet (in this case Minnie) who watches the stories alongside the subjects. At the end of the story, the puppet is asked to say what she thinks happened in the story. In our version, instead of asking subjects if the puppet is 'right' or 'wrong' (as in the original TVJT), we asked whether the puppet 'answered well' (i.e., *Apantise kala;,* 'Did-(she)-answer well?'). This modification was made since we were interested in felicity, not truth. Finally, the subjects were asked to justify their answers by explaining why they thought that Minnie answered well or not.

The children were tested individually in a quiet room away from the class. Adult subjects were shown a videotaped version of the stories witnessed by the children, including the warm-up stories. They were given a score sheet and were instructed to indicate, for each story, whether Minnie had 'answered well' or not. They were also asked to provide a brief justification for their answers.

For each scale, subjects were asked to judge four statements like the ones in (9-11) - see Table 1. As the reader can verify from the Table, the critical stories were identical in the case of *some/all* and *two/three*. In addition to the critical statements in Table 1, and for each scale, subjects were also asked to judge four control statements like the ones in (15) and (16) - see Table 2:

(15) The Smurf bought two of the rings/balloons.

(16) The Smurf bought some of the rings/balloons.

The purpose of these controls was to ensure that subjects, and in particular children, could accept or reject the puppet's statements when appropriate and, more importantly, that they could do so when these statements involved felicitous uses of terms like *some* and *two*. For each control statement, the experimenter had a choice between two versions: one that

was a correct description of the story and would therefore elicit a 'Yes' response and one that was an incorrect description of the story and would therefore elicit a 'No' response. The experimenter selected the version of the control statement (correct or incorrect description) based on the child's response of the preceding critical statement. If the child had rejected the puppet's statement on the previous critical trial, the experimenter selected the version of the control statement that would elicit a 'Yes' response, and vice-versa. This step was taken to keep a balance between 'Yes' and 'No' responses. Finally, each child received two 'warm-up' stories, one designed to elicit a 'Yes' answer and the other a 'No' answer. The puppet's statements on the warm-up stories are also given in Table 1.

INSERT TABLE 1 & 2

Subjects (5-year-olds and adults) were randomly assigned to one of three conditions, determined by scale type (i.e., *some/all*; *two/three*; *start/finish*) which gave rise to a 2X3 design with age and scale type as between subject factors and 10 subjects per cell (Table 3). The age range and mean ages for the 10 children assigned to each scale condition, i.e. *some/all*, *two/three* and *start/finish* are 5;0 to 5;11 (mean 5;4), 4;11 to 5;10 (mean, 5;3) and 5;0 to 5;11 (mean, 5;4) respectively. In each condition, subjects received four critical trials and four control trials administered in a pseudo-random order. Within each condition, order of presentation was counterbalanced between subjects.

INSERT TABLE 3

Results

In the analysis below, our dependent measure is the proportion of 'No' responses to the puppet's statements, i.e. the subjects' tendency to judge these statements as 'bad' descriptions of the stories they witnessed. Beginning with test trials, we found that adult subjects overwhelmingly rejected the puppet's statements in each of the three conditions, i.e.

92.5% of the time in the *some/all* condition, 100% of the time in the *two/three* condition and 92.5% of the time in the *start/finish* condition. Statistical analysis revealed no reliable difference between these rejection rates ($F(2, 27) = 1.92$, $p = 0.16$). By contrast, we found that while 5-year-olds rejected the puppet's statements in the case of *two/three* 65% of the time[7], they almost never did so in the case of *some/all* and *start/finish* (12.5% and 10% of the time respectively). This difference was confirmed statistically ($F(2 ,27) = 11.17$, $p < 0.001$). Pairwise comparisons (Tukey Kramer) further revealed a reliable difference between *two/three* - *some/all* and *two/three* - *start/finish* ($p = 0.002$ and $p = 0.001$ respectively) but no reliable difference between *some/all* and *start/finish* ($p = 0.77$). The proportions of 'No' responses were entered into an analysis of variance (ANOVA) with two factors: age (5-year-olds vs. adults) and scale type (*some/all*, *two/three*, *start/finish*). The analysis revealed a main effect of age ($F(1,54) = 135.34$, $p < 0.0001$), a main effect of scale type ($F(2,54) = 13.03$, $p < 0.0001$) and a reliable interaction between age and scale type ($F(2,54) = 7.43$, $p = 0.001$) (see Figure 3).

INSERT FIGURE 3

On the control items, adults gave correct responses 100% of the time in the *some/all* condition, 80% of the time in the *two/three* condition and 95% of the time in the *start/finish* condition. No reliable difference was found between these means ($F(2,27) = 2.43$, $p=0.1$). On the same items, children gave correct responses 90% of the time in the *some/all* condition, 95% of the time in the *two/three* condition and 85% of the time in the *start/finish* condition. Here again, no reliable differences between the means were found ($F(2,27) =1.28$, $p = 0.29$).

Recall that in addition to accepting or rejecting the puppet's statements, subjects were asked to provide justifications for their answers. Adults, who overwhelmingly rejected the

[7] Specifically, 5 children rejected the puppet's statements on all 4 of the test trials, 1 child on 3 of the test trials, 1 child on 2 of the test trials, 1 child on 1 of the test trials and 2 children rejected all 4 of the test trials. In sum, 6 of the 10 children almost always rejected the puppet's statements (i.e. on 3 or 4 of the test trials), 3 children almost never rejected the puppet's statements (i.e. on either 0 or 1 of the test trials) and one child rejected half of the test trials and accepted the other half.

puppet's statements on all three scales, typically justified their negative answers by invoking statements containing the stronger terms on each of the respective scales (i.e., *all*, *three* and *finish*). So for example, when the puppet described a situation in which each of the three horses in the story jumped over the fence by saying that some of the horses jumped over the fence, adult subjects said that the puppet was wrong because ALL of the horses jumped over the fence. Similarly, when each of the three horses jumped over the fence and the puppet described the situation by saying that two of the horses jumped over the fence, adults explained that the puppet was wrong because TWO of the horses jumped over the fence. Finally, when the puppet described a situation in which a girl had completed a puzzle by saying that the girl had started making the puzzle, adults objected on the grounds that the girl had in fact FINISHED making the puzzle. Justifications of this kind, making direct and explicit reference to the stronger terms on each of the respective scales (i.e., *all* instead of *some*, *three* instead of *two,* and finish instead of *start*) accounted for 98% of the adult subjects' justifications.

Unlike adults, 5-year-olds massively accepted the puppet's statements on the *some/all* and *start/finish* scales, i.e. 87.5% and 90% of the time respectively. Here, children's justifications fall into two main categories: a) justifications involving a replica of the puppet's statement, and b) justifications invoking the stronger term on the respective scales. So for example, upon hearing the puppet describe a situation in which all of the horses jumped over the fence by saying that some of the horses jumped over the fence, children would typically explain that the puppet was right either "because some of the horses jumped over the fence" (type a), or "because all of the horses jumped over the fence" (type b). Children overwhelmingly produced type (a) justifications. In the case of *some/all*, 74% of children's justifications were of type (a) and 22% were of type (b) (the remaining 4% were irrelevant justifications). In the case of *start/finish*, 55% of the children's justifications were of type (a) and 36% were of type (b) (the remaining 9% were irrelevant justifications). On the *two/three* scale, children rejected the puppet's statements much more often than on the other two scales (i.e. 65% vs. 12.5% and 10%). Here, when children rejected the puppet's statements, they always invoked the fact that three - not two - of the characters had performed the relevant action. When children incorrectly accepted the puppet's statements (i.e. 35% of the time), 42% of their justifications were of type (a) and 35% of type (b) (the remaining 20% (i.e. 3 out of 14) were irrelevant justifications).

Discussion

The first, important conclusion that can be drawn from the results presented above concerns the psychological reality of scalar implicatures. According to pragmatic theory, the use of scalar terms such as *some*, *two* and *start* in sentences like *Some of the horses jumped over the fence*, *Two of the horses jumped over the fence,* and *The girl started making the puzzle* invites the inference that not all of the horses jumped over the fence, no more than two horses jumped over the fence, and the girl didn't finish making the puzzle respectively. In accordance with these predictions, we found that adults overwhelmingly rejected statements containing *some*, *two* and *start* (i.e., 92.5% of the time, 100% of the time and 92.5% of the time respectively) in situations that satisfied the truth conditions of stronger terms on the corresponding scales, i.e. *all*, *three* and *finish*. Moreover, adults typically justified their rejections of the puppet statements precisely by invoking statements containing the stronger terms on the relevant scales (i.e. 98% of the time). This finding provides clear evidence that scalar implicatures are indeed generated during language comprehension (see Noveck, 2001 for a similar conclusion).

A second, striking observation is the apparent lack of sensitivity displayed by 5-year-olds vis-à-vis the implicatures associated with the interpretation of the terms *some* and *start*. Unlike adults, children rejected statements containing these terms in situations satisfying the truth conditions of stronger terms on the respective scales only 12.5% of the time in the case of *some* and 10% in the case of *start*. In other words, children accepted the puppet's statements 87.5% and 90% of the time respectively. One possible interpretation of this finding is that children have a general tendency to accept anything the puppet says. However, children's responses to the control items (i.e., 90% correct responses for *some*/*all* and 85% correct responses for *start*/*finish*) clearly show that this cannot be the case. Recall that the control items were specifically designed to ensure that children are capable of giving both kinds of answers, i.e. YES and NO. That is, every time a child answered YES to the puppet statement on a critical trial, the following control item was designed to elicit a NO answer and vice-versa. In this case, children's near perfect response patterns on the control items

demonstrates not only that they did not experience any difficulty with the task itself but, crucially, that they are capable of giving both YES and NO answers.[8]

One of our motivations in choosing scalar terms like *some*, *two* and *start* was to test children's sensitivity to scalar implicatures on the basis of simple cases. Recall from our earlier discussion that previous attempts typically relied on more complex cases (e.g., modals, Novek 2001; interaction of *or* with universal quantifier, Chierchia et al. 2001).[9] It is therefore possible that previous reports of children's apparent lack of sensitivity to scalar implicatures were due in part to the complexity of the exemplars chosen by these investigators. The results presented here, however, not only comport with previous observations but, more importantly, appear to suggest that children fail to derive pragmatic interpretations even in the simplest cases.

A third noteworthy observation emerging from the results presented above is the fact that children's apparent lack of sensitivity to scalar implicatures does not generalize to all scalar terms. Surprisingly, while children only rejected the puppet's statements 12.5% and 10% of the time in the case of *some/all* and *start/finish*, they did so significantly more often in the case of *two/three*, i.e. 65% of the time ($p < .01$). Incidentally, this result lends further credence to the conclusion that children's high acceptance rate in the case of *some/all* and *start/finish* is not due to a general YES bias. Recall that our motivation for choosing to investigate children's interpretation of numeral terms stems from their controversial theoretical status. While investigators may disagree about the precise semantic and pragmatic status of such terms, there are reasons to believe that numerals are different from other scalar terms (see section 2). It is therefore interesting that this theoretically motivated difference between numerals and other scalar terms is reflected in children's behavior, as shown by their differential treatment of the terms under investigation. We come back to this observation and consider its consequences in the General Discussion section.

There is now substantial evidence that preschoolers, unlike adults, appear not to be sensitive to the pragmatic interpretation of scalar terms. Experiment 1 provides additional evidence supporting this conclusion. Why then do preschool children, who are otherwise linguistically sophisticated, seem to be oblivious to interpretations that are so natural for

[8] Other work using the same technique has shown that children of the same age (and even younger children) can - and routinely do - give both answers (Musolino, Crain and Thornton 2000; Lidz and Musolino submitted).

[9] On the difficulties associated with the acquisition of modal expressions, see Papafragou (1998, 2000).

adult speakers? One possible answer is that preschoolers are utterly unable to derive such interpretations. Another possibility – one that hasn't been systematically explored so far – is that children's non-adult behavior has to do with the nature of the task that is being used to investigate the status of their pragmatic abilities. All of the experiments discussed so far (including our own) have used some version of the Truth Value Judgment Task. This task has proven to be an extremely useful tool in the investigation of children's syntactic and semantic knowledge. However, since it was not originally designed to elicit pragmatic judgements, its use in testing pragmatic understanding may somehow mask early communicative abilities.

The assumption underlying the use of this method in assessing a comprehender's sensitivity to conversational implicatures has been that the hearer would be able to realize, without being directly instructed to do so, that the purpose of the task is not to determine whether a given sentence is true or false in a given context but rather whether the sentence in question can be used felicitously in that context. In our version of the TVJT, we tried to direct participants' attention to felicity rather than truth by asking whether Minnie 'answered well' rather than whether she was right or wrong. Success in our task presupposes the ability to infer the experimenter's goal and interpret this broad question accordingly. As it turns out, adult speakers seem to be quite good at doing that. For example, upon being presented with a statement that is semantically true in a given context adults seem to be able to readily infer that, since the sentence is true in that particular context, the experimenter must be asking something else, namely whether the sentence in question is felicitous in that particular context. It is less clear that children would be as consistent in drawing such inferences.[10]

A related concern involves the contexts in which the scalar implicatures were expected to arise. Recall that scalar inferences do not appear indiscriminately every time a term such as *some* or *start* is used but their derivation is crucially constrained by expectations about intended cognitive effects. Adults, reasoning flexibly about the purpose of our task, can 'see through' to the informativeness and relevance expectations raised by our scenarios. For instance, they can judge that the use of *some* is inappropriate in a context which licenses the

[10] One further reason to suspect that this may be so is that our warm-ups did not distinguish between truth and felicity in an optimal way. For instance, when Minnie describes a log as a table, children may well reject that statement on the basis that it is simply false, rather than infelicitous. Even though we attempted to make this statement not simply false, but outrageously false, it may certainly have biased children to interpret the task as a test for truth/falsehood (rather than felicity/infelicity).

use of *all*. If preschoolers, unlike adults, cannot readily infer the pragmatic nature of the task, *and are not given adequate motivation to go beyond the truth conditional content of the utterance*, they may readily settle for a statement which is true but does not satisfy the adult expectations of relevance and informativeness.[11]

In sum, these observations raise the possibility that the difference between children and adults noted so far may lie not in their respective abilities to compute scalar implicatures but rather in their ability to infer the goal of the experiment. In other words, it is conceivable that children take our task at face value, as it were, and - in the absence of additional motivation - provide answers on the basis of whether a statement is semantically true in a given context. This scenario would be perfectly compatible with our results (as well as those of previous studies). Moreover, this approach makes the interesting prediction that it should be possible to boost children's pragmatic interpretations of statements containing scalar terms, to the extent that the presentation conditions of the task (i.e. scenarios and instructions) make them aware of the goals of the experiment. This prediction is addressed and tested in experiment 2.

Experiment 2

Experiment 2 preserves the basic design of experiment 1 in that subjects are asked to judge statements containing the scalar terms *some*, *two* and *start* in situations which satisfy the truth conditions of stronger terms on the respective scales (i.e., *all*, *three* and *finish*). However, three important modifications were made in order to test the hypothesis that children's apparent inability to derive scalar implicatures may be due to the nature of the task and in particular children's inability to infer the goals of the experimenter. First, we decided to enhance children's awareness of the goal of the task by initially training them to detect pragmatic anomaly. Second, the stories witnessed by our subjects were modified so as to create a context in which the main character's performance (subsequently described by the puppet) becomes the focal point of the story. Finally, instead of being asked to describe what happened in the story, the puppet is asked to directly comment on the main character's performance. Each of these manipulations is described in more detail in the sections below.

[11] The developmental literature has repeatedly stressed the importance of motivation and context in children's performance on a variety of cognitive tasks. We discuss this point fully in the General Discussion section.

Overall, the intended effects of these manipulations were (a) to enhance the child's awareness of the goals of the task (judging felicity vs. truth), and (b) to present contexts which readily invite the kind of pragmatic inferences that the task tries to elicit.

Method

Subjects

30 Greek-speaking children (11 boys and 19 girls) ranging in age between 5;1 and 6;5 (mean 5;7) participated in this experiment. These children were recruited at daycare centers in the same Athens area as the children used in experiment 1.

Procedure

In an initial training phase, children we presented with four warm-up stories designed to enhance their awareness of the fact that they were being asked to produce pragmatic judgments. Children were first told that the puppet, Minnie, sometimes said 'silly things' and that the purpose of the game was to help Minnie to 'say things better'. For example, Minnie would be shown a toy dog which she would describe using the truth-conditionally accurate - but pragmatically infelicitous - statement 'This is a little animal with four legs'. The child would then be asked whether 'Minnie answered well' and whether 'we can say it better'. In case the child failed to correct the puppet and provide a better description, the experimenter eventually corrected Minnie and provided the appropriate description, i.e. 'Minnie didn't say that very well. This is a DOG'. Two of the four warm-up statements were examples like the one just described (involving the use of an accurate but infelicitous statement) and the other two were cases where Minnie would provide a description that was both accurate and felicitous. For example, Minnie would be shown a toy elephant and would describe it by saying 'This is an elephant'. This step was taken as a precaution to ensure that children wouldn't think that Minnie always said silly things. The complete set of warm-ups is shown in Table 4.

INSERT TABLE 4

As before, subjects witnessed four test stories in which they were asked to judge statements containing the scalar terms *some*, *two* and *start* in situations which satisfied the truth conditions of the stronger terms of the respective scales, i.e. *all*, *three* and *finish*. However, the stories in Experiment 2 were all based on scenarios in which the main character was involved in a contest or a challenge. The main character's performance therefore became the focal point of the stories and at the end, the puppet was asked to comment on how well the character in question had done, 'How did X do?' (*Pos ta pige o X?).* In one of the stories for example, one of the characters claims that he is very good at throwing hoops around a pole and he challenges Mickey to try and do the same with three hoops. Mickey really concentrates hard and he's able to put all the hoops around the pole. At the end of the story, Minnie is asked "How did Mickey do?" and she answers by saying that "Mickey put some of the hoops around the pole". The idea behind this manipulation is to raise the expectations of cognitive effects so that only an answer making reference to *all* the hoops would satisfy the demands of the communicative situation: given that Mickey's performance is being directly evaluated, it certainly matters whether he puts only some or all of the hoops around the pole. We expected that such contexts would be more conducive to providing the kinds of pragmatic inferences under investigation.

Materials

As described above, children were initially trained to detect pragmatic anomaly on the basis of the statements shown in Table 4. Children then heard four test stories and four control stories administered in a pseudo-random order. As before, order of presentation was counterbalanced between subjects within a condition. The puppet's statements on the test and control items are given in Tables 5 and 6 respectively. It is to be noted that the manipulation described above regarding the test stories also applied to the control stories.

INSERT TABLES 5 & 6

Finally, as in Experiment 1, subjects were randomly assigned to either of the three scale conditions (i.e. *some/all*, *two/three*, *start/finish*) yielding the design in Table 7 where scale

condition was treated as a between subject factor. The age range and mean ages for the 10 children assigned to each scale condition, i.e. *some/all*, *two/three* and *start/finish* are 5;1 to 6;2 (mean 5;6), 5;4 to 6;3 (mean 5;7) and 5;5 to 6;5 (mean 5;9), respectively.

INSERT TABLE 7

Results

As before our dependent measure was children's *Yes/No* responses to the puppet's statements. Recall that in Experiment 1, children rejected the puppet's statements 12.5% of the time in the *some/all* condition, 10% of the time in the *start/finish* condition and 65% of the time in the *two/three* condition. What we found here is that the manipulations described above led children to reject the puppet's statements much more often, i.e. 52.5% of the time for the *some/all* scale, 47.5% of the time for *start/finish* [12] and 90% of the time for *two/three*. We compared the rejection rates from Experiment 1 and Experiment 2 by entering them into a 2 (training vs. no training) by 3 (*some/all, two/three, start/finish*) ANOVA. The analysis revealed a main effect of training ($F(1,54) = 14.61$, $p = 0.0003$), a main effect of scale type ($F(2,54) = 12.28$, $p < 0.0001$) and no reliable interaction between training and scale type ($F(2,54) = 0.26$, $p = 0.76$) (see Figure 4). On control items, children gave correct responses 85% of the time in the *some/all* condition and 95% of the time in the *two/three* and the *start/finish* condition. No reliable differences between these means were found ($F(2,27) = 0.73$, $p = .48$).

INSERT FIGURE 4

[12] Specifically, on the *some/all* scale, 4 children rejected the puppet's statements on all 4 target trials, 1 child rejected on 3 trials, 4 children never rejected and 1 child rejected on half of the trials. In sum, 5 children almost always rejected the puppet's statements (i.e. on either 3 or 4 of the test trials), 4 children almost never rejected the puppet's statements (i.e., on either 0 or 1 of the test trials) and 1 child was equivocal. On the start/finish scale, 5 children almost always rejected the puppet's statements (i.e., 3 children rejected on all 4 trials and two children on 3 trials) and the other 5 children almost never rejected the puppet's statements (i.e., 4 children never rejected the puppet's statements and 1 child rejected on 1 of the 4 trials).

Having established that children reject the puppet statements reliably more often as a result of the manipulations described above[13], we now turn to the justifications that children provided when they rejected the puppet's statements. Recall from Experiment 1 that adults, who overwhelmingly rejected the puppet's statements on all three scales, typically justified their negative answers by invoking statements containing the stronger terms on each of the respective scales (i.e., *all*, *three* and *finish*). So for example, when the puppet described a situation in which each of the three horses in the story jumped over a fence by saying that some of the horses jumped over the fence, adults subjects said that the puppet was wrong because ALL of the horses jumped over the fence. Also recall that justifications of this kind, making direct and explicit reference to the stronger terms in each of the respective scales (i.e., *all* instead of *some*, *three* instead of *two* and *finish* instead of *start*) accounted for 98% of the adult subjects' justifications. What we found here is that the justifications provided by children following their negative responses to the puppet's statement were exactly of the same kind as those provided by adults in Experiment 1. In other words, children in experiment 2 typically rejected the puppet's statements by invoking statements containing the stronger terms on each of the respective scales. This type of response accounted for 93% of children's justifications in cases where children rejected the puppet's description.

Discussion

The results of Experiment 2 suggest that children's sensitivity to scalar implicature greatly improves once they are made aware of the goals of the task and provided with contexts which more readily invite the kinds of pragmatic inferences under investigation. In interpreting the results of Experiment 2, it is crucial to observe that children's tendency to correct the puppet did not extend to the control items (as witnessed by children's near-perfect responses on those items). In other words, while children did correct the puppet

[13] Note that the children in experiment 2 are on average slightly older than those in experiment 1 (i.e. mean 5;3 vs. 5;7, $p < .05$). This is because some of the children in experiment 2 were tested later during the school year. However, there are strong reasons to believe that the effect observed in experiment 2 was not due to this small difference in age. First, in at least one of the three conditions, i.e. *some/all*, there is no reliable difference in age between the group of children in experiment 1 and 2 (mean 5;4 vs. 5;6, respectively, $t(18) = 0.985$, $p= .33$). In this case therefore age cannot be responsible for children's improved performance. Second, and more importantly, previous studies (see section 3) indicate that children well beyond the age of 5 and up to the age of 10 are insensitive to scalar implicatures. Given this fact, it is extremely unlikely that small differences in age between 5-year-olds have any impact on their ability to spontaneously derive scalar implicatures.

when she described a situation in which Mickey had put all the hoops around the pole by saying that he had put some of the hoops around the pole, they did not correct the puppet when she described a situation in which a little girl jumped over a fence by saying "The little girl jumped over the fence". In this case, children never corrected the puppet by saying, for example, "A little girl WITH A RED DRESS jumped over the fence". This is important because it shows that children's improved performance on the test items must come from a more refined appreciation of scalar inference and not from a general tendency to correct the puppet regardless of what she says. Finally, it is important to point out that while the manipulations described above did have a noticeable effect on the child's ability to derive scalar inference, our data do not allow us to determine the respective contribution of each of these manipulations (e.g. the role of training vs. the change in extralinguistic context). Further experimentation would be required in order to tease apart the role of these factors.

6. General discussion

The experiments presented here explore young children's ability to derive pragmatic inferences during utterance comprehension. Our specific target was a well-known group of pragmatic inferences, *scalar implicatures*, which arise whenever a 'weaker' proposition (e.g. *Some men have beards*) is used to communicate that a stronger proposition does not hold (*Not all men have beards*). We were mainly interested in whether young (but otherwise linguistically sophisticated) children compute such implicatures, as adult communicators routinely do. We also explored the role of (a) the semantics of specific scalar expressions, and (b) the context of the experimental scenarios, as well as the nature of the task, in the derivation of these conversational inferences.

Using data from Modern Greek, our experiments tested three different scales, *meriki/oli* ('some/all'), *dio/tris* ('two/three') and *arxizo/teliono* ('start/finish'). In Experiment 1, subjects were presented with contexts which satisfied the truth conditions of the stronger (i.e. more informative) terms on each scale (i.e., *all*, *three* and *finish*) but were described using the weaker terms of the scales (i.e., *some*, *two*, *start*). The results showed that, while adults overwhelmingly rejected these infelicitous descriptions, 5-year-old children almost never did so; children also differed from adults in that their rejection rate on the numerical scale was reliably higher than on the two other scales. Experiment 2 made a number of significant changes to the

design of Experiment 1: First, through both pre-test training and more specific instructions, it was strongly brought to the children's attention that their task was to detect pragmatic infelicity (rather than judge truth). Second, the stories were modified so that they now supported the implicatures in a much stronger way. These manipulations gave rise to significantly higher rejection rates on the part of the children than those observed in Experiment 1. Interestingly, even though children generally still fell short of fully mature performance, in the number scale they now performed in an adult-like way.

These findings have several implications for the psychology of conversational implicature. Overall, they establish that scalar implicatures are psychologically real for adult interlocutors, thereby confirming the strong intuition that such inferences are regularly derived during mature communication. Furthermore, the Greek data support the assumption that conversational inferences of this sort are universal and arise whenever certain informativeness and relevance prerequisites are met at the semantics-pragmatics interface. From a developmental perspective, our results confirm previous experimental findings which showed that children are not as sophisticated as adults in attending to these subtle aspects of the semantics-pragmatics interface. More importantly, they provide clear empirical evidence for both the role of semantics and the effects of context and task characteristics on children's derivation of scalar implicature. As we now show, if semantics and context conspire in appropriate ways, children are much better at spontaneously deriving scalar implicatures.

One important conclusion which emerges from this work is that children's success in computing scalar inferences within a single 'scale' depends on their awareness of the goals and of the general nature of the task. This conclusion agrees with a large body of work which emphasizes the role of context and of verbal instructions in children's performance on a variety of reasoning tasks. For instance, it is well known that children who have failed classic Piagetian tasks can succeed if the experimental conditions become simpler and more child-friendly and/or the child is made aware of the purpose of the experimental questions (see Bever, Mehler and Epstein 1968, Gelman and Greeno 1989, Greeno, Riley and Gelman 1984, Shipley 1979, Markman and Seibert 1976, Rose and Blank 1974, Samuel and Bryant 1984, McGarrigle and Donaldson 1974).[14] Related work has underlined that, with enough

[14] Similar conclusions have been reached in research on adult reasoning, where the importance of participants' interpretations of a psychological task, and the influence of pragmatic factors on this interpretation, are

motivation and contextual support, even young children may succeed in computing other people's mental states, paying attention to intentionality cues and dealing with pragmatic inferences in a variety of tasks and communicative situations (Siegal, Waters and Dinwiddy 1988, Cassidy 1998, Hurewitz, Trueswell, Gleitman and Brown-Schmidt 2000, Papafragou in press).

Even if the goals of the experimental 'game' are clear, there are several reasons why children may approach the communicative situation differently from adults - and hence fail to derive scalar implicatures. One possible explanation is that children may have a different assessment of the communicative expectations raised by the experimental set-up. Recall the story in which a group of horses ends up jumping over the fence. For adults, it is infelicitous to describe the event by saying *Some of the horses jumped over the fence*, since in fact all of them did. Five-year-old children, however, may not have a similar threshold for expected relevant information. If so, they may accept a weaker statement instead of the stronger one, which is exactly what they do. Another possible explanation is that children find it too effort-demanding to entertain both alternatives, the stronger and the weaker one, and to choose between them (a prerequisite for scalar implicature).[15] These explanations may work jointly in accounting for children's early difficulties with scalar inference. For instance, if combined, they predict that, if children are provided with a context where communicative expectations are clear and where the stronger alternative to the weaker statement is made particularly salient, children will be more prone to noticing the implicature. Our second experiment, which introduces clear cognitive expectations (which always involve assumptions containing the stronger term of the scale), shows that this prediction is borne out.

There is some independent evidence which suggests that such processing factors (expectations of cognitive gains, effort in processing alternatives) may be responsible for children's early difficulties with scalar implicatures. In the first place, young communicators are not completely oblivious to expectations of relevant information. As anyone who has spent five minutes with a child knows, children do not talk like little logicians, producing *some* where *all* would have been appropriate (and true), or asserting a possibility when certainty is warranted. This shows that, in production at least, even preschoolers have the

---

increasingly recognized (Politzer and Noveck 1991, Evans 1995, Sperber, Cara and Girotto 1995, Thompson 2000, Van der Henst, Politzer and Sperber in press).

[15] These explanations are inspired by discussions of the role of cognitive expectations and processing effort in utterance interpretation by Sperber and Wilson (1985/1995); cf. also Noveck (2001).

ability to differentiate between the informational affordances of the different members of a scale. A similar conclusion can be reached using some facts about comprehension. Recall that Chierchia et al. (2001) found that, given two alternatives which differ in informational strength, children prefer the strongest one if it is justified by the experimental scenario. Again this shows that, when the pressures of computing a scalar inference on-line are removed, children show some preference for the strongest term warranted by relevance. What is novel about our research is that it shows that, given the right kind of contextual support, children show some success in *spontaneously* deriving scalar implicatures. Furthermore, our method leaves open the possibility that other manipulations of cognitive effects and the salience of the 'stronger' alternatives could yield even higher success rates on the part of young participants.

Naturally, despite their improved performance in Experiment 2, children's pragmatic skills are much less robust than adults' - and this explains why this difference persists even in favorable contexts. This conclusion is hardly surprising. Much current and older developmental work shows that preschoolers' ability to use, co-ordinate and assess multiple (semantic and contextual) cues is fragile and unstable (Trueswell et al. 1999, Markman and Seibert 1976, Gelman and Greeno 1989, Shatz 1978, Ackerman 1983). More generally, children have been reported to have problems in a variety of pragmatic tasks involving the computation of implicatures (Shatz 1980). Taken together with this literature, our present work suggests that the ability to compute conversational inferences cannot be viewed as an all-or-nothing affair. Rather, developmental research should pay attention to the conditions under which children's pragmatic success or failure is observed in the hope of uncovering how pragmatic mechanisms develop. Furthermore, given the importance of task demands on children's pragmatic performance, the research reported here has specific methodological implications for the design of tests tapping into early communication skills.

We finally come to what seems to us one of the most intriguing aspects of our findings - namely, the conclusion that number terms seem to behave very differently from the other two scalar terms in our studies. The contrast between *some/all* vs. *two/three* is especially remarkable given that our stories and materials were identical in these two cases. There are several reasons that may be involved in children's higher success rates with number terms. One is the possible interference from the counting routine. Children of this age love to engage in counting and very often used counting in our experiments in order to justify their

responses. For instance, when Minnie offered *Two of the horses jumped over the fence*, several children protested by saying *No, one, two, THREE horses jumped over the fence* (while at the same time counting the horses one by one). Explicit counting of this sort may affect the computation of the scalar inferences in a number of ways. It may offer a specific and precise way of verifying statements containing number terms (by placing the referents of the corresponding NPs in a one to one correspondence with objects in the world). Counting games may also encourage an 'exact' interpretation of the numerals. Notice that neither of these steps is available for *some/all*. The fact that, in our experimental scenarios, *two* is therefore more clearly 'anchored' than the more vague *some* (or, similarly, *start*), may be responsible for the higher success rates it reveals on the part of the children.

These observations raise interesting issues about the theoretical status of the numerical scale and its relation to the rest of the scalar predicates. As we mentioned in Section 2, there are several reasons to consider numerical scales distinct from regular scales, and our experimental data seem to confirm this difference. However, the exact implications for linguistic theory, especially the semantics-pragmatics of numerals, are less clear at this stage. What seems obvious is that, in child language at least, cardinals do not seem to have an 'at least' semantics: if they did, then children in Experiment 1 should not have rejected weakly informative statements such as *Two of the horses jumped over the fence*, since such statements would be true on the 'at least' reading of *two*. In general, the most plausible (even though not the only) conclusion we can draw from our present results is that children assign either an 'exact' or an underspecified semantics to the numerals. This goes against the standard ('lower-bounded') view of the semantics of numerals but conforms with more recent semantic proposals about number terms, according to which their semantic content is underspecified between an 'exactly', an 'at least' and an 'at most' reading (Carston 1990, 1998, Horn 1992).

Notice that, if both children and adults can be shown to have an underspecified semantics for numerals, this would lend support to continuity assumptions, according to which children and adults share the same semantic representations. But if it turns out, for independent reasons, that children have an 'exactly' semantics and adults an underspecified semantics for numerals, then one would have to accept the presence of a fundamental discontinuity; the next task would then be to show how children can revise their semantic entry for cardinals so as to arrive at the adult semantics. It is worth pointing out that, in the

considerable developmental literature which looks at children's acquisition of number terms (Carey 2001, Gelman and Gallistel 1978, Gelman 1993, Wynn 1992, Bloom and Wynn 1997 among many others), it is usually assumed that both children and adults have an 'exact' semantics for number terms. Moreover, according to one influential position, children assign meaning to cardinal expressions in natural language by placing them in a one-to-one correspondence with an innate conceptual 'integer list' (Gelman and Gallistel 1978). Even though this literature looks at much younger children than we did, it would be interesting to see whether older children indeed have an 'exactly' semantics for the numerals. The linguistic and developmental theories of number seem to be a good case where semantic-pragmatic theory and acquisition research should mutually inform and constrain each other (for a similar conclusion, see Papafragou 1998, 2000).

There are several questions raised by the present work on scalar implicatures. Our goal in this paper was to compare different ways in which semantics and pragmatics contribute to the computation of these conversational inferences. An important task for future work will be to use these developmental findings to test predictions of different pragmatic models and to tease apart in a more detailed way the respective contributions of fine aspects of the semantics-pragmatics interface to the calculation of such inferences.

Acknowledgements

Thanks to Lila Gleitman, Henry Gleitman, John Trueswell, and the members of the CHEESE seminar at the University of Pennsylvania for comments and suggestions. We wish to thank the teachers and children at the 3rd and 5th daycares at Vrilissia (Athens), Greece. We are also indebted to Professors D. Chila-Markopoulou, D. Theofanopoulou-Kontou and S. Hoidas for their help with the experiments conducted at the University of Athens.

References

Ackerman B. (1983). Children's judgements of the functional acceptability of referential communications in discourse contexts. *Journal of Child Language* 10: 151-166.

Begg I. and G. Harris (1982). On the interpretation of syllogisms. *Journal of Verbal Learning and Verbal Behavior* 21: 595-620.

Bever T., J. Mehler and J. Epstein (1968). What children do in spite of what they know. *Science* 162(3856): 921-924.

Bloom P. and K. Wynn (1997). Linguistic cues in the acquisition of number words. *Journal of Child Language* 24/3: 511-533.

Braine M. and B. Rumain (1981). Children's comprehension of "or": evidence for a sequence of competencies. *Journal of Experimental Child Psychology* 31: 46-70.

Cacciari C. and S. Glucksberg (1994). Understanding figurative language. In M.A. Gernsbacher (ed.), *Handbook of Psycholinguistics*, 447-477. San Diego, CA: Academic Press.

Carey S. (2001). Cognitive foundations of arithmetic: Evolution and ontogenesis. *Mind and Language* 16/1: 37-55.

Carston R. (1988). Implicature, explicature, and truth-theoretic semantics. In R. Kempson (ed.), *Mental representations: The interface between language and reality*, 155-181. Cambridge: Cambridge University Press.

Carston R. (1990). Quanity maxims and generalised implicature. *UCL Working Papers in Linguistics* 2: 1-31. Reprinted in *Lingua* 96 (1995): 213-244.

Carston R. (1998). Informativeness, relevance, and scalar implicature. In R. Carston and S. Uchida (eds.), *Relevance theory: Applications and implications*. Amsterdam: Benjamins.

Cassidy K. (1998). Three- and four-year-old children's ability to use desire- and belief-based reasoning. *Cognition* 66: B1-B11.

Chierchia G., S. Crain, M. T. Guasti, A. Gualmini and L. Meroni (2001). The acquisition of disjunction: Evidence for a grammatical view of scalar implicatures. In A. H.-J. Do et al (eds.), *BUCLD 25 Proceedings*, 157-168. Somerville, MA: Cascadilla Press.

Clark H. (1992). *Arenas of language use*. Chicago: University of Chicago Press.

Crain S. and R. Thornton (1998). *Investigations in Universal Grammar: A guide to research on the acquisition of syntax and semantics*. Cambridge, MA: MIT Press.

Evans J. (1995). Relevance and reasoning. In S. E. Newstead and J. Evans (eds.), *Perspectives on thinking and reasoning*. Hove, UK: Laurence Erlbaum.

Fauconnier G. (1975). Pragmatic scales and logical structure. *Linguistic Inquiry* 6: 353-375.

Gazdar G. (1979). *Pragmatics*. New York: Academic Press.

Gelman R. (1993). A rational-constructivist account of early learning about numbers and objects. In D. Medin (ed.), *Learning and motivation*, Vol. 30, 61-96. New York: Academic Press.

Gelman R. and J. Greeno (1989). On the nature of competence: principles for understanding a domain. In L. Resnick (ed)., *Knowing and learning: Essays in honor of Robert Glaser*, 125-186. Hillsdale, NJ: Erlbaum.

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

Gibbs R. (1994). *The poetics of mind*. Cambridge: Cambridge University Press.

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

Greeno, J., M. Riley and R. Gelman (1984). Conceptual competence and children's counting. *Cognitive Psychology* 16: 94-134.

Gualmini A., S. Crain, L. Meroni, G. Chierchia and M. T. Guasti (2001). At the Semantics/Pragmatics Interface in Child Language. *Proceedings of SALT XI*. Cornell University, Ithaca, NY.

Harnish R. (1976). Logical form and implicature. In T. Bever, J. Katz and T. Langedoen (eds.), *An integrated theory of linguistic ability*, 313-391. New York: Crowell.

Hirschberg J. (1985). *A theory of scalar implicature*. Doctoral diss., University of Pennsylvania.

Horn L. (1972). *On the semantic properties of the logical operators in English*. Doctoral diss. UCLA Distributed by IULC, Indiana University.

Horn L. (1992). The said and the unsaid. In C. Barker & D. Dowty (eds.), *Proceedings of SALT II*, 163-192. Dept. of Linguistics, Ohio State University.

Hurewitz F., J. Trueswell, L. Gleitman and S. Brown-Schmidt (2000). The walrus and the parser: Developing the ability to use contextual and lexical information for syntactic ambiguity resolution. Paper presented at the 14th Annual CUNY Conference on Human Sentence Processing, University of Pennsylvania, 15-17 March.

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

Lidz J. and J. Musolino (submitted). Children's command of quantification.

Markman E. (1981). Comprehension monitoring. In P. Dickson (ed.), *Children's oral communication skills*. New York: Academic Press.

Markman E. and J. Seibert (1976). Classes and collections: Internal organization and resulting holistic properties. *Cognitive Psychology* 38: 561-577.

McGarrigle J. and M. Donaldson (1975). Conservation accidents. *Cognition* 3: 341-350.

Mill J.S. (1867). *An examination of Sir William Hamilton's philosophy*. 3rd ed. London: Longman.

Monck W. H. S. (1881). *Sir William Hamilton*. London: Sampson, Low.

Musolino J., S. Crain and R. Thornton (2000). Navigating negative quantificational space. *Linguistics* 38/1: 1-32.

Musolino J. and J. Lidz (forthcoming). Preschool logic: Truth and felicity in the acquisition of quantification. To appear in *Proceedings of BUCLD 26*.

Newstead S. E. (1995). Gricean implicatures and syllogistic reasoning. *Journal of Memory and Language* 34: 644-664.

Newstead S. E. and R. A. Griggs (1983). Drawing inferences from quantified statements: a study of the square of opposition. *Journal of Verbal Learning and Verbal Behavior* 22: 535-546.

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

Papafragou A. (1998). The acquisition of modality: Implications for theories of semantic representation. *Mind and Language* 13/3: 370-399.

Papafragou A. (2000). *Modality: Issues in the Semantics-Pragmatics Interface*. Amsterdam/New York: Elsevier Science.

Papafragou A. (in press). Mindreading and verbal communication. To appear in *Mind and Language* 17/1&2: 55-67.

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

Politzer G. and I. Noveck (1991). Are conjunction rule violations the result of conversational rule violations? *Journal of Psycholinguistic Research* 20: 83-103.

Rose S. and M. Blank (1974). The potency of context in children's cognition: An illustration through conservation. *Child Development* 45: 499-502.

Sadock J. (1984). Whither radical pragmatics? In D. Shiffrin (ed.), *Georgetown University Round Table on Language and Linguistics*, 139-149. Washington, DC: Georgetown University Press.

Samuel J. & P. Bryant (1984). Asking only one question in the conservation experiment. *Journal of Child Psychology and Psychiatry* 25: 315-318.

Shatz M. (1978). The relationship between cognitive processes and the development of communication skills. In C. Keasey (ed.), *Nebraska Symposium on Motivation*, 1-42. Lincoln: University of Nebraska Press.

Shatz M. (1980). Communication. In P. Mussen (ed.), *Handbook of Child Psychology, Vol. 3: Cognitive Development* (vol. eds. J. Flavell & E. Markman), 841-889. New York: Wiley.

Shipley E. (1979). The class inclusion task: Question form and distributive comparison. *Journal of Psycholinguistic Research* 8: 301-331.

Siegal M., L. Waters and L. Dinwiddy (1988). Misleading children: Causal attributions for inconsistency under repeated questioning. *Journal of Experimental Child Psychology* 45: 438-456.

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

Sperber D., F. Cara and V. Girotto (1995). Relevance theory explains the selection task. *Cognition* 57: 31-95.

Sperber D. and D. Wilson (1995). *Relevance: Communication and cognition*. Cambridge, MA: Harvard University Press. 2nd ed. 1995.

Thompson V. (2000). The task specific nature of domain-general reasoning. *Cognition* 76: 209-268.

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

Van der Henst J.-B., G. Politzer and D. Sperber (in press). When is a conclusion worth deriving? A relevance-based analysis of indeterminate relational problems. To appear in *Thinking and Reasoning*.

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

Figures and Tables

Figure 1

The horses are about to jump over the fence

Figure 2

Some of the horses jumped over the fence

Figure 3

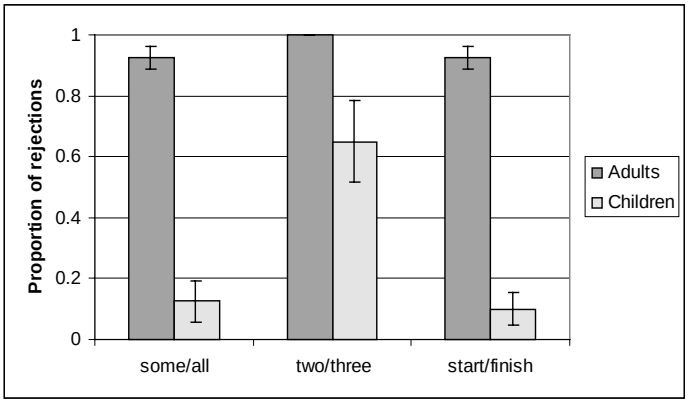

Figure 4

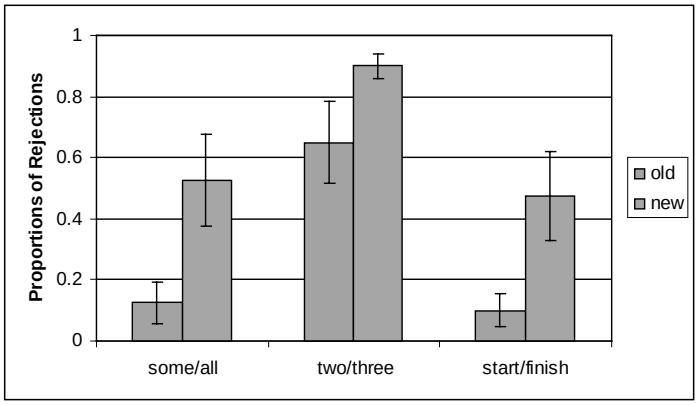

Table 1

Scale type	Puppet statements on warm-ups and critical trials
Warm-ups	This is a horse (True: subjects are shown a horse)
	This is a table (False: subjects are shown a log)
Critical trials	
Some/all	Some of the horses jumped over the log
	Some of the rabbits went on the merry-go-round
	Some of the dinosaurs ate trees
	Some of the Smurfs bought dogs
Two/three	Two of the horses jumped over the log
	Two of the rabbits went on the merry-go-round
	Two of the dinosaurs ate trees
	Two of the Smurfs bought dogs
Start/finish	The Smurf started painting the balloons
	The Smurf started putting the cars into the bag
	The little girl started making the puzzle
	The circus man started bringing down the elephants

Table 2

Scale type	Puppet statements on control trials
Some/all	The Smurf bought two of the rings/balloons The karate guy broke two of the boards/cars Donald found two of the bad guys/animals Donald cleaned two of the airplanes/cars
Two/three	The Smurf bought some of the rings/balloons The karate guy broke some of the boards/cars Donald found some of the bad guys/animals Donald cleaned some of the airplanes/cars
Start/finish	The Smurf bought two of the rings/balloons The karate guy broke some of the boards/cars Donald found some of the bad guys/animals Donald cleaned two of the planes/cars

Table 3

	Some/all	Two/three	Start/finish
5-year-olds	n=10	n=10	n=10
Adults	n=10	n=10	n=10

Table 4

Puppet statements on warm-up stories	
Felicitous statements	
This is a house	(pointing to a house)
This is an elephant	(pointing to an elephant)
Infelicitous statements	
This is a little animal with four legs	(pointing to a dog)
These are branches	(pointing to a tree)

Table 5

Scale type	**Puppet statements on critical trials**
Some/all	Mickey put some of the hoops around the pole The little tiger lifter some of the blocks The big guy caught some of the horses The bunny put some of the pieces in the puzzle
Two/three	Mickey put two of the hoops around the pole The little tiger lifter two of the blocks The big guy caught two of the horses The bunny put two of the pieces in the puzzle
Start/finish	Mickey started putting the hoops around the pole The little tiger started making the tower Donald started coloring the star The bunny started making the puzzle

Table 6

Scale type	**Puppet statements on control trials**
All scales	The little girl jumped over the fence The giraffe ate the apple The red frog caught the calamari The horse beat the turtle

Table 7

	Some/all	Two/three	Start/finish
5-year-olds (no training)	n=10	n=10	n=10
5-year-olds (training)	n=10	n=10	n=10