LANGUAGE AND COGNITIVE PROCESSES
2005, 20 (5), 667–696

Psychology Press
Taylor & Francis Group

Why children and adults sometimes (but not always) compute implicatures

Maria Teresa Guasti and Gennaro Chierchia
Università di Milano-Bicocca, Milan, Italy

Stephen Crain
University of Maryland at College Park, MD, USA

Francesca Foppolo
Università di Milano-Bicocca, Milan, Italy

Andrea Gualmini
Massachusetts Institute of Technology, Cambridge, MA, USA

Luisa Meroni
University of Maryland at College Park, MD, USA

Noveck (2001) argued that children even as old as 11 do not reliably endorse a scalar interpretation of weak scalar terms (*some*, *might, or*) (cf. Braine & Rumain, 1981; Smith, 1980). More recent studies suggest, however, that children's apparent failures may depend on the experimental demands (Papafragou & Musolino, 2003). Although previous studies involved children of different ages as well as different tasks, and are thus not directly comparable, nevertheless a common finding is that children do not seem to derive scalar implicatures to the same extent as adults do. The present article describes a series of experiments that were conducted with Italian speaking subjects (children and adults), focusing mainly on the scalar term *some*. Our goal was to carefully examine the specific conditions that allow the computation of implicatures by children. In so doing, we demonstrate that children as young as 7 (the youngest age of the children who participated in the Noveck study) are able to compute implicatures in experimental conditions that properly satisfy certain contextual prerequisites for deriving

Correspondence should be addressed to Università deglí Studi di Milano-Bicocca, Dipartimento di psicologia, P.zza dell'Ateneo Nuovo 1, 20126 Milano, Italia. Email: mariateresa.guasti@unimib.it

http://www.tandf.co.uk/journals/pp/01690965.html DOI: 10.1080/01690960444000250

such implicatures. We also present further results that have general consequences for the research methodology employed in this area of study. Our research indicates that certain tasks mask children's understanding of scalar terms, not only including the task used by Noveck, but also tasks that employ certain explicit instructions, such as the training task used by Papafragou & Musolino (2003). Our findings indicate further that, although explicit training apparently improves children's ability to draw implicatures, children nevertheless fail to achieve adult levels of performance for most scalar terms even in such tasks, and that the effects of instruction do not last beyond the training session itself for most children. Another relevant finding of the present study is that some of the manipulations of the experimental context have an effect on all subjects, whereas others produce effects on just a subset of children. Individual differences of this kind may have been concealed in previous research because performance by individual subjects was not reported. Our general conclusions are that even young children (7-year olds) have the prerequisites for deriving scalar implicatures, although these abilities are revealed only when the conversational background is natural.

INTRODUCTION

Recent studies on the development of scalar implicatures suggest that otherwise linguistically competent children treat *some* as compatible with *all* and accept under-informative statements such as *some giraffes have long necks* much more often than adults do. In this respect, they appear to be 'more logical than adults' because they disregard the pragmatic norms that lead adults to reject under-informative statements. In this article, we investigate the extent to which this conclusion is correct by manipulating the experimental tasks. We provide evidence of the emergence of specific aspects of pragmatic competence in children. In addition, we discuss conditions that might favour or inhibit the use of children's pragmatic competence.

In ordinary conversational exchanges, as well as in reasoning tasks conducted in the laboratory, many factors influence the way subjects respond. In concrete conversational settings, our understanding of language is influenced not only by the propositional content of a statement, i.e., by truth conditions, but also by pragmatic factors. Generally speaking, these pragmatic factors are responsible for certain inferences that are drawn on the basis of assessing what is said against the background of what could have been said. One widely discussed example of pragmatic inferences in language use is scalar implicatures. Scalar implicatures are components of a message that are not directly encoded in the statement uttered by a speaker, but arise from aspects of the conversational dynamics. Following Grice (1975) and much literature inspired by him, it is argued that if a speaker says (1a), the hearer is entitled to assume that the speaker intended to convey (1b).

(1) a. Some students passed the exam
b. Some student passed the exam, *but not all did*

The added statement 'but not all did' is not part of the propositional content of (1a); rather, it is an inference that is invited by the speaker's decision to use *some* instead of other quantified expressions that might have been used in its place. The added piece of information 'but not all did' is a scalar implicature.[1] We say that (1b) represents the scalar component or the pragmatically enriched meaning of *some* whereby *some* comes to mean *some but not all*. This can be contrasted with the logical meaning of *some* according to which it means *some and possibly all*. We interpret (1) the way we do, because logical words, like the quantifier *some*, form a scale with other expressions (i.e., *many*, *most*, *every*). Uttering (1) activates the following statements that could have been uttered in its place (see Horn, 1972):

(2) a. Every student passed the exam
b. Most students passed the exam
c. Many students passed the exam

Notice that the statements in (2) are naturally ordered in terms of their informativeness, with the statement (2a) being the strongest member of the scale. In fact, the *every* statement in (2a) entails all the other statements below it and, similarly, (2b) entails (2c). That is, if every student passed the exam then it surely is the case that most students did, that many students did and that some students did. Standard Conversational norms invite the illocutionary agent to reason as follows (Grice, 1975). If the speaker chooses to produce the statement *some students passed the exam*, it must be because either he has no evidence that the stronger statements hold or, perhaps, he knows that the stronger statements do not hold. So, assuming that the speaker is well informed and being cooperative, the hearer will tend to infer that the stronger statements do not hold and, upon hearing (1a), she will assume that *not all students passed the exam* is an intended consequence of the speaker's assertion.

According to the dominating views of the semantic-pragmatic interface, scalar implicatures are calculated globally, within a pragmatic module,

[1] Scalar implicatures are cancellable. In fact, we can continue (1a) with the denial of the implicature without incurring in a contradiction, as seen in (i):

(i) Some students passed the exam. In fact, they all did.

In this respect, scalar implicatures differ from entailments, which give rise to contradiction if they are negated.

(ii) Some students passed the exam. In fact, none of them did.

after the semantic module has assigned recursively the truth conditions to the sentence (see Levinson, 2000 for discussion). This view is challenged in Chierchia (2004) on the grounds that there are embedded scalar implicatures, a fact that is unexpected on the globalist approach. Consider the utterance in (3a) and let us concentrate on the implicature triggered by the second disjunct. Intuition tells us that (3a) implicates (3b).

(3) a. Mary is either reading a paper or seeing some students
b. Mary is either reading a paper or seeing some, though not all, students

However, if implicatures were calculated globally, it is not obvious how the implicature in (3b) would come about. According to the globalist view, the alternative relative to the second disjunct of (3a) could be (4a). Observe that (4a) is stronger and hence more informative than (3a). By choosing (3a), the speaker implicates that all stronger alternatives are denied. Thus, in particular (4a) is denied, which amounts to (4b). But (4b) entails (4c), which contradicts what is stated in (3a) and this is unwanted.

(4) a. Mary is either reading a paper or seeing **every** student
b. Mary is either reading a paper or seeing some students and it is **not the case** that Mary is reading a paper or seeing every student
c. Mary is not reading a paper

This problem is known as 'the disjunction problem'. On the basis of examples such as these, along with considerations about other linguistic phenomena, Chierchia, Crain, Guasti, Gualmini, and Meroni (2001) propose to abandon the globalist view in favour of an approach whereby implicatures are factored into semantic representations locally, by a recursive mechanism parallel to the standard one that derives the logical meaning of sentences. So, on Chierchia's proposal, implicatures are computed incrementally while meaning is derived and not after semantics has delivered the logical meaning of sentences to the pragmatic module. Under this view, the logical and scalar meanings of statements are not clearly distinct, since implicatures are integrated during the course of their interpretation. As a consequence of the architecture of the language apparatus, we are not led to expect children to be less competent than adults in deriving the scalar meaning than in deriving the logical meaning. Given the assumption that both logical meaning and implicatures are computed within the semantic module, observed differences between adults and children are more likely to arise because the derivation of the scalar meaning adds to the complexity of language processing by consuming additional processing resources. In light of these proposals

about how implicatures are derived and about the consequences of this for language acquisition, we now turn to a consideration of developmental data.

Developmental studies have shown that children initially tend to interpret scalar terms logically, where older children and adults show more sensitivity to pragmatic implicatures. For example, Smith (1980) found that 4- to 7-year-old children who are competent with quantifiers treat *some* as meaning *some and possibly all* in a task in which they had to answer questions like *Do some giraffes have long necks?* Similarly, Braine and Rumain (1981) established that 7- to 9-year-olds tend to interpret *or* inclusively, i.e., they take *A or B* to mean *A or B and possibly both.* In a recent paper, Noveck (2001) argues that, despite adult-like competence in dealing with many logical statements, children between the ages of 7 and 11 years old differ from adults in that they do not readily access the scalar meaning of weak scalar terms (e.g., *some* and *might*); children stick instead to the logical meaning, while adults tend to assign the scalar meaning.

One interpretation of the experimental findings reported in Noveck (2001), and in previous work, is that 7- and 11-year-old children do not derive scalar implicatures or, more generally, are incapable of pragmatic inferencing. A second possibility, proposed explicitly by Noveck, is that 'pragmatic interpretations become evident subsequent to logical interpretations' because their derivation involves a cost. In the experiment carried out by Noveck, the cost may actually be due to the materials used: the statements that were employed demanded subjects to draw upon real world knowledge. It seems important to establish whether or not this is so, as differences in real world knowledge are not the only possible source of the observed differences between children and adults. Using different materials and a different task, one that involved more explicit instructions, Papafragou and Musolino (2003) attempted to see if the experimental findings of the Noveck study could be an artifact of the particular task that was utilized. Papafragou and Musolino (2003) found, as Noveck had found before them, that children were not as sensitive as adults to the implicatures that are associated with *some*. However, their performance improved when children received training that was meant to make them aware of the need to give a pragmatic judgement.

The results of the studies by Noveck (2001) and Papafragou and Musolino (2003) cannot be directly compared, since the subjects were of different ages, different methods were used and different questions were posed to the subjects. Complicating the issue further is the observation that the pattern of responses by adult subjects also differed across the two studies. Adult subjects in the Noveck study failed to compute the implicature associated with *some* 41% of the time. By contrast, adult subjects in the Papafragou and Musolino study computed the relevant

implicatures 93% of the time; a similar finding was reported in Chierchia et al. (2001). So, the Noveck study indicates that adults, too, fail to compute implicatures in certain conditions. Regardless, the results of these two studies cannot help us settle the question of whether it was real world knowledge that matters in children's failure to derive implicatures in the Noveck study, since Papafragou and Musolino did not use materials that required subjects to consult real-world knowledge, but they found nevertheless that children were less prone to derive scalar implicatures than adults. Similarly, these studies leave open the possibility that children are simply incapable of pragmatic inferencing until age 11. This is quite surprising, but it is buttressed by the fact that children failed to spontaneously derive implicatures in both the Noveck study and in the Papafragou and Musolino study. Finally, the results do not allow us to conclude whether the problem in inferring implicatures depends on cognitive factors or on linguistic factors. Since Piaget's work and research that has ensued from it, it is well documented that context can deeply influence the emergence of a cognitive capacity which is potentially available to children. At the same time, it is possible that children may not perform a linguistic task because they fail to have the relevant linguistic knowledge. To address these issues, we attempted to study the development of pragmatic inferences and the factors that influence these inferences in children.

All in all, the findings of the previous literature have led to two hypotheses about why children may not derive scalar implicatures. One possibility is that young children simply lack the prerequisites for deriving scalar implicatures, a problem that may persist for several years. Let us call this the Pragmatic Delay hypothesis. The findings of previous research are open to an alternative interpretation, however. They could be interpreted as showing that children can, in principle, compute scalar implicatures, but that they do so to a lesser extent than adults in contexts that impose demands on processing resources. We will call this the Pragmatic Limitation hypothesis.

In this paper, we scrutinize both of these hypotheses in a series of experimental studies. Our goals are twofold. On the one hand, the previous experimental findings invite us to investigate the factors that give rise to variation in the behaviour of both children and adults and this, in turn, invites us to reassess the materials and methods used in previous research to measure the pragmatic abilities of both children and adults. In so doing, we point out some facts that have been overlooked in previous research concerning the patterns of responses by individual subjects, both by adults and by children. On the other hand, we aim to show that 7-year-olds derive implicatures to the same degree as adults when the contexts meet all the cognitive and linguistic requirements for doing so. We believe

that this is a result against which to compare further research on the development of pragmatic abilities. We limit our study to the investigation of children's interpretation of the scalar term *some*, but we do not anticipate any differences in children's or adults' understanding of other scalar terms. Similarly, we chose 7-year-old children to make sure that our results could be compared with those obtained by Noveck. If 7-year-olds can compute implicatures as well as adults do under conditions that reduce excessive demands on their computational resources, we expect that older children can do so as well. It is possible that children younger than 7 will remain less able to compute implicatures, as work by Papafragou and Musolino seems to suggest. However, before tackling this issue, it needs to be established that 7-year-olds, at least, have reached adult-like competence in inferring implicatures. This will provide a baseline against which to compare further work to help disentangle cognitive and linguistic factors that might be involved in the derivation of pragmatic inferencing.

The paper is structured as follows. First, we partially replicate Noveck's experiment with Italian speaking subjects (adults and 7-year-old children) to establish whether there is a developmental pattern. This experiment serves as a baseline for further investigation. Next, we manipulate the experimental situation by adding a training session to Noveck's experimental task, much in the spirit of Papafragou and Musolino (2003). The training component of the study is designed to establish whether children's performance improves when the experimental goals are clarified. We then discuss the interpretations proposed by Noveck (2001) and Papafragou and Musolino (2003). Finally, we proceed to report a task where the youngest subjects tested by Noveck (7-year-olds) are able to compute scalar implicatures at the same level as adults. This is achieved in a study using the Truth Value Judgment Task (Crain & Thornton, 1998). Although this task does not provide explicit instructions to children, it does make available to children all of the relevant evidence for the derivation of implicatures. We argue, therefore, that the failure of 7-year-old children to compute scalar implicatures in the Noveck study probably resulted from the absence of explicit contextual support, and not from children's inability to compute scalar implicatures. If children lacked the competence to compute implicatures altogether, then alterations in the context are not expected to evoke adult-like performance. We conclude with a few remarks about the pragmatic abilities of children younger than those who participated in the present studies.

EXPERIMENT 1: REPLICATION OF NOVECK (2001)

The first experiment is a partial replication of the experiment conducted by Noveck (2001), who assessed the ability of French-speaking children's

comprehension of sentences including the quantifiers *some* and *all.* Noveck presented 7- to 11-year-old children and adults with a series of sentences with the form *Some X [verb]* and *All X [verb]*. Some of these sentences were true, others were false and some of them were logically true, but pragmatically infelicitous (*Some giraffes have long necks*). Subjects' task was to answer whether or not they agreed with the statements. Noveck found that there was a significant difference between the two groups of children and the adults with respect to pragmatically infelicitous sentences. Adults tended to reject these statements much more than the two groups of children, rejection rate being 41% for adults, 89% for 7-year-olds, and 85% for 11-year-olds. With respect to other statements, children and adults behaved in the same way, namely they rejected false statements and accepted true statements.

The present experiment serves as a baseline for the remaining experiments. It differs from the study by Noveck in three respects. First, it was conducted with Italian-speaking children, and second, it includes only children at the youngest age tested in the Noveck study, namely 7-year-olds. Limiting the age of child subjects in this way is justified for two reasons. For one thing, the pattern of responses by the 11-year-old children did not differ from that of the 7-year-olds on the critical statements in the Noveck study. The second reason is based on our aim to determine the conditions in which children at some age can draw implicatures as much as adults; success in achieving this aim with younger children will sustain a generalisation to older children. Third, Noveck had two lists of statements, while we used only one list. The two lists differed in that the same content (e.g., giraffes have long necks) was presented with either the quantifier *some* or with the quantifier *all,* with each participant hearing only one kind of statement. We did not include this manipulation, first because it did not produce any effect in Noveck's experiment and, second, because it would have complicated the design of the other experiments.

Methods

Participants. Eighteen 7-year-olds (age range 7;0–7;6 years, mean age 7;2) and nineteen adult native speakers of Italian participated in the experiment. Children were recruited from the primary school of Cernusco sul Naviglio, near Milan. Adults were volunteers from the University of Milano Bicocca, Department of Psychology.

Materials and design. The materials were essentially the same as those employed by Noveck (2001) with some modifications. They were presented in Italian, with some changes in lexical items. There were 15 sentences with *some* and 15 with *all*, based on three types of information: factually

universal (that birds have wings is best expressed with the quantifier *all*), factually existential (that birds live in cages is best expressed with the quantifier *some*) and absurd (that stories are made of bubbles is false with either quantifier). In this way, for each quantifier, there were three sets of six different statements, as summarised below:

(a) five absurd *some* sentences (e.g., some stories are made of bubbles)
(b) five true (and felicitous) *some* sentences (e.g., some children are blond)
(c) five true (but pragmatically under-informative) *some* sentences (e.g., some giraffes have long necks)
(d) five absurd *all* sentences (e.g., all doors sing)
(e) five true *all* sentences (e.g., all birds have wings)
(f) five false *all* sentences (e.g., all birds live in cages)

Each participant was presented with the entire randomised sequence of *some* sentences (15 in all), followed by the entire sequence of *all* sentences, or vice versa. The English translation of the test materials is presented in Appendix A.

Procedures. The task was a statement evaluation task. Participants were told that they were going to listen to a number of statements that had been given to the experimenter by a friend and that they were asked to say whether they agreed with each statement or not. They were also told that, if they did not agree with a statement, they would occasionally be asked to explain why, so that the experimenter could inform her friend about why such statements were inaccurate.

Results and discussion

The main finding of Experiment 1 is that children accept statements like *some giraffes have long necks* much more often than adults do: 87% compared with 50%. This essentially replicates the finding of the Noveck study of French-speaking children. With respect to all the other statements, children are highly competent, as can be seen in Table 1.

As in Noveck, we submitted the data to ANOVA, using a 2 (age: children, adults) $\times$ 2 (order of presentation of quantifiers) $\times$ 6 type of statements ((a) through (f) above)) design, with percentages of logically correct responses serving as the dependent measure. The analysis revealed an effect of age, $F_1(1, 33) = 5.891$, $p < .05$ and an effect of type of statement, $F_1(5, 165) = 16.82$, $p < .0001$; in addition, there was an interaction between age and type of statement, $F_1(5, 165) = 7{,}66$, $p < .0001$. Before examining the main effects, it pays to analyse the interaction.

TABLE 1
Rates of correct responses to the six types of sentences presented in Experiment 1: Replication of Noveck (2001)

Sentence type	*Correct response*	*Children*	*Adults*
All-statements			
Absurd (*All windows talk*)	No	95	98
Appropriate (*All pots have handles*)	Yes	98	97
Inappropriate (*All dogs are black and white*)	No	97	100
Some-statements			
Absurd (*Some dogs speak French*)	No	97	94
Appropriate (*Some tulips are yellow*)	Yes	95	98
Inappropriate (true but pragmatically infelicitous) (*Some cats have hair*)	No	87	50

It turned out that the interaction was essentially due to one statement type, the under-informative statements (e.g., *some giraffes have long necks*). Thus, the main effects were due to this kind of statement. For this reason, we limited further analyses to this statement type. A 2 (age: children, adults) $\times$ 2 (order of presentation: *Some*-statements, *All*-statements) ANOVA with percentages of logically correct responses as a dependent measure reveals a main effect of age, $F_1(1, 33) = 8.73$, $p < .005$ with children accepting the relevant statement ($M = .877$ $SD = .29$) more than adults did ($M = .505$, $SD = .449$). We carried out the analysis of the other five kinds of statements, but did not find any significant effect. This confirms that the main effects in the global analysis were due to the under-informative statements containing *some.* Finally, since the distribution was highly asymmetrical, we transformed the data into z-scores (using the square root of the arcsine) and repeated the analyses. However, this did not yield any differences from the previous analysis. We also entered the data into an analysis by items with age and type of statements as factors, using percentages of logically correct responses as the dependent measure. We found a main effect of age, $F_2(1, 48) = 52.14$, $p < .00001$, a main effect of types of statements $F_2(5, 48) = 152.76$, $p < .00001$ and an interaction between age and type of statements, $F_2(5, 48) = 70.2$, $p < .00001$. This interaction was essentially due to one kind of statement, the under-informative statements introduced by *some*, whose rate of acceptance by children was higher than that of adults. An ANOVA by items limited to

the statements judged by children revealed a significant effect of statement type, $F_2(5, 24) = 8.33$, $p < .001$. A post-hoc Scheffé test indicates that this effect was due to the under-informative statements, which differed from all the other five statement types ($p < .05$). The same effect is revealed in an items analysis on the statements produced by adults, $F_2(5, 24) = 239.92$, $p < .0001$; the post-hoc Scheffé test revealed that the under-informative statements differed from all the others ($p < .0001$). We conducted an ANOVA by subjects (bearing in mind that the data do not yield a normal distribution); Figure 1 displays the distribution of subjects as a function of the number of times they accepted the critical statements (from 0 up to 5 times).

Figure 1a shows that the vast majority of the children never rejected the target statements. Figure 1b, instead, shows that adults either (almost) always accepted the target sentences, as children, or (almost) always rejected them, thereby yielding a bimodal distribution.[2] To establish whether the developmental effects brought out by the ANOVA were reliable, subjects were divided in two groups: one group was comprised of subjects that accepted the critical statements 3 or more times, and the other group was comprised of subjects who accepted the critical statements less than 3 times. It turned out that 89% of the children accepted the critical statements 3 or more times, whereas these statements were accepted by adults just 47% of the time. An analysis of proportion was applied to the data, yielding a significant result ($p = .01$). Thus, both the ANOVA and the analysis of proportion indicate that there is a clear developmental trend, with more children than adults accepting under-informative statements like *some giraffes have long necks*.

Children's responses on other logical statements indicate a high degree of competence: they rejected the absurd *some* and *all* statements and, when asked to explain their rejection, they responded appropriately, saying, e.g., that birds don't have telephones or that people do not eat books. They accepted or rejected statements with both of these quantifiers in felicitous conditions. Thus, statements like *some cakes are made of chocolate, all birds have wings* were generally accepted by children and *all birds live in cages* were generally rejected and the explanation was that there are birds that are not in cages, but free. In the few cases in which children rejected under-informative sentences like *Some giraffes have long necks,* they motivated their answer by pointing out that all giraffes have long necks. The explanations that children offered were similar to those offered by adults, indicating that they evaluated these statements in the same way as adults did. That is, they evaluated whether a given statement corresponded

[2] Adults' responses gave rise to a bimodal distribution also in Noveck's study, but no attention was devoted to this fact.

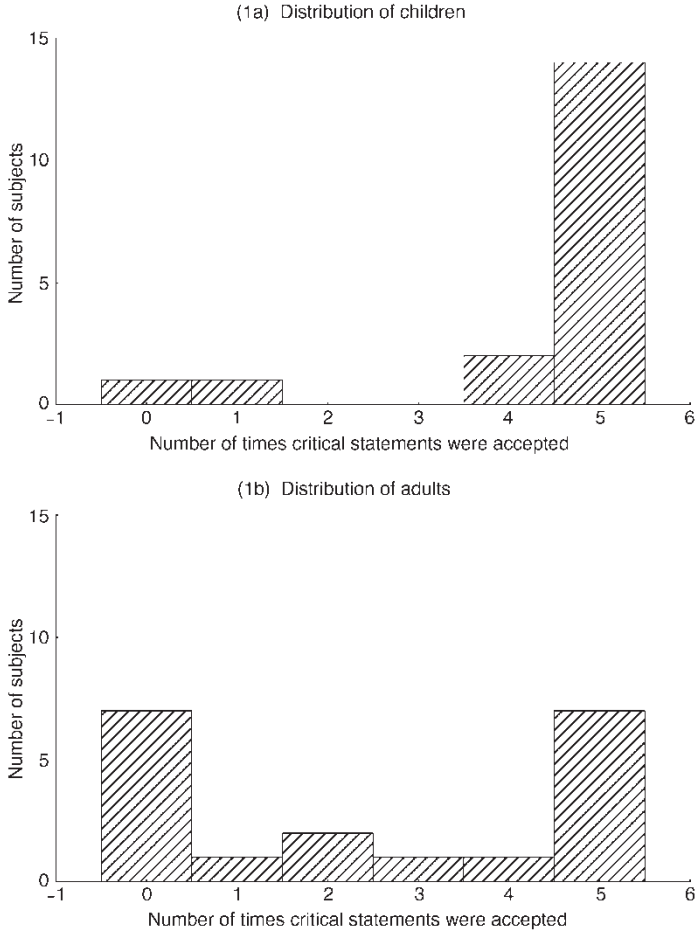

Figure 1. Distribution of children (1a) and adults (1b) depending on the number of times they accepted critical statements in the replication of Noveck's (2001) experiment.

or not to a general state of affairs in the world, using a binary judgement of truth value. At first glance, the results of this experiment confirm Noveck's conclusion that implicatures are not evident in children and that, consequently, the logical meanings of *some* and *all* take priority.

As we already mentioned, one can offer two possible explanations for these findings: children could lack the ability to draw implicatures (Pragmatic Delay hypothesis) or they could have this ability, but are prevented from displaying their pragmatic competence by some feature of the experimental design (Pragmatic Limitation hypothesis). Several aspects of the experiment suggest that this second possibility is worth pursuing. For one thing, although the level of performance by adults was

far superior to that of children, adult performance on the under-informative statements was far from optimal. Adults rejected the test sentences only about half of the time. One reason for this, and also for children's poor performance, perhaps, was that the instructions were unconstrained. On the one hand, subjects were asked to evaluate statements without evidence that was directly available; on the other hand, they were simply asked whether they agreed with the statements or not. It was assumed that subjects would base their judgements on the facts that obtain in the real world and would therefore behave as they would in ordinary exchanges. However, subjects were free to adopt other strategies, including conjuring up other contexts, as the basis for the evaluation of the test statements. Thus, a potential experimental confound is the absence of any explicit context. The influence of the absence of context is investigated in an experiment reported in section 4, using the Truth Value Judgement Task.

Another potential source of children's poor performance is that children did not understand the experimental instructions, as suggested in a work by Papafragou and Musolino (2003). Children may have failed to appreciate that they were being asked to evaluate the informativeness of statements and not simply their truth or falsity. To investigate this issue, we conducted a second experiment adopting similar procedures to those used in the Papafragou and Musolino (2003) study.

EXPERIMENT 2: MANIPULATION OF EXPERIMENTAL DEMANDS

In Experiment 1, children apparently displayed a high percentage of 'logical' responses. According to Papafragou and Musolino (2003), this may be due to the fact that they have taken the experimental task to be about the truth or falsity of the statements, and not about their informativeness or lack thereof. In this experiment, we attempted to enhance this second interpretation of the experimental instructions, by having children participate in a training session before the test phase, as in Papafragou and Musolino.

Methods

Participants. Twenty-one Italian-speaking children (age range 7;0–7;7 years, mean age: 7;1) recruited from the primary school of Cernusco sul Naviglio (Lombardia).

Materials and design. This experiment used the same material and design as in Experiment 1.

Procedures. Unlike in Experiment 1, initially children participated in a training session consisting in the presentation of four figures depicting a grape, a cook, a cake, and a chair. The experimenter introduced the figures to the child by saying that they were given by a friend who asked for the child's help. For each figure the friend had indicated two ways of describing it and wanted to know from the child which way was better. Each description was a true description of the object, but one was more specific than the other. For the grape, it was said that the friend used the terms *grape* and *fruit* to describe the object; similarly, the terms *cook* and *man* were used to describe the cook, *cake* and *sweet stuff* were used for the cake, and *chair* and *piece of furniture* were used for the chair. It was decided, in advance, that children who erred on two out of four trials would not be invited to continue with the testing phase of the experiment. At the end of the training session, the experimenter reminded the child that there are different ways of describing an object, as seen in the training session, and sometimes there is one way of describing objects/events that is better than another way. Then, as in Experiment 1, it was said to the child that she was going to listen to a series of statements and would be asked to say whether she agreed or not. She was also told that if she did not agree, she would occasionally be invited to explain why.

Results and discussion

In the training session, all children had no hesitation in choosing the most restrictive term to describe the relevant object, e.g., *grape* rather than *fruit.* Therefore, all children continued with the experiment. The main result of the test phase of the experiment is that children who participated in the training session rejected statements like *Some giraffes have long necks* to a much greater extent than did children who were not trained: the rejection rate was 12% for children without training, and it rose to 52% with training. To establish whether training significantly affects performance, children's correct responses from Experiment 1 were compared with the rate of logically correct responses in the present experiment. An ANOVA was carried out, with percentages of logically correct responses as a dependent measure, using the following design: 2 (condition: without training, with training) $\times$ 2 (order of presentation of quantifiers) $\times$ 6 (type of statements). There was a main effect of condition, $F_1(1, 35) = 7.08$, $p < .01$ and of type of statements, $F_1(5, 175) = 19.48$ $p < .00001$; in addition, an interaction was found: between condition and type of statements, $F_1(5, 175) = 9.49$, $p < .00001$. As in the previous experiment, the interaction was essentially due to one kind of statement, i.e., the under-informative statements including *some.* For this reason we limited further analysis to just this kind of statement. A 2 (condition: without training,

with training) × 2 (order of presentation of quantifiers) ANOVA with percentages of logically correct responses as a dependent measure reveals a main effect of condition, $F_1(1, 35) = 10.88$, $p < .005$ with children in Experiment 1 agreeing with under-informative statements ($M = .877$, $SD = .29$) more than in the present experiment ($M = .466$, $SD = .43$). An analysis by items on the material used in this experiment reveals a significant effect of type of statements, $F_2(5, 24) = 326.87$, $p < .0001$; post-hoc Scheffé test shows that the effect is due to the under-informative statement that differs from all the others ($p < .001$). Training has an effect. However, as in the previous experiment, the use of the mean as a measure of central tendency is not entirely telling, since there is a high degree of dispersion in the data. Figure 2 displays the distribution of the child subjects as a function of the number of times they accepted the under-informative statements (from 0 up to 5 times). Unlike Experiment 1 (see Figure 1a), the figure makes it clear that the children' s responses formed a bimodal distribution, such that subjects either always disagreed or always agreed with the critical statements. In short, training had a strong effect on some children, but no effect whatsoever on other children.

To establish whether the results of the ANOVA were reliable, we divided subjects into two groups. One group included subjects who accepted the critical statements three or more times, and the other group was comprised of subjects who accepted these statements less than three times. Of the children who were trained, 48% accepted the critical under-informative statements on three or more trials, while this percentage rose to 89% for children without training. The data were subjected to an

Figure 2. Distribution of children depending on the number of times they accepted critical statements in Experiment 2. Children had received a training session before being tested.

analysis of proportion, which showed that the effect of training was significant ($p = .01$). Interestingly, after training, children behaved exactly as our adult subjects in Experiment 1 (compare Figure 2 and Figure 1b). In both cases, there was a bimodal distribution, with children either always accepting or always rejecting the critical statements. We conclude that training enhances the rejection of under-informative statements in 7-year-olds. It is important to keep in mind that training does not have an effect on all children, but when it does, children consistently rejected the critical statements and explained their rejections by invoking a more informative description; for example, they disagreed with *Some giraffes have long necks*, and justified this response by explaining that *all giraffes have long necks*. Thus, some children as young as those in the youngest group tested by Noveck were able to derive the scalar meaning of *some* when their awareness of the criteria to be used in making responses in the experimental task was enhanced by training.

Two questions arise. The first question concerns the persistence of the effect of training on children's performance. Do children that have been trained maintain the same level of performance when they are retested a period of time after the first test, without a new training session prior to the retest? We turn to this question in Experiment 3. The second question concerns the optimal level of performance by children overall. Although some children clearly benefited from training, others did not. Earlier we pointed to another factor that could have deflated children's performance, the absence of context. It is conceivable that all 7-year-olds will benefit from the use of context, regardless of training. This possibility will be examined in experiment 4, using the Truth Value Judgment Task.

EXPERIMENT 3: DOES THE EFFECT OF TRAINING PERSIST?

Methods

Experiment 3 investigates whether the enhancement of performance achieved through training is permanent. We assess this possibility by retesting the same children who participated in Experiment 2 one week after the first test, without repeating the preliminary training session.

Subjects. The subjects were the same as those who took part in Experiment 2.

Materials and design. The materials were modeled after those used in Experiment 2, but the content of the statements was changed. The complete list of sentences is given in Appendix B.

Procedures. The same procedure were employed as in Experiment 1, that is, children were not provided with a training session.

Results and discussion

The main finding of this experiment is that children who rejected the critical *some* statements after training (Experiment 2) failed to do so when retested without additional training. The rate of rejection dropped from 52% in Experiment 2 to 22% in the present experiment. To establish the effect of training on subsequent behaviour (without additional training), we compared children's correct responses in Experiment 2 to those in the present experiment. We submitted the data to ANOVA, with percentage of correct responses as a dependent measure, using a design with the following factors: 2 (condition: test with training, retest without training) $\times$ 2 (order of presentation of quantifiers) $\times$ 6 (type of statements). A main effect of type of statement was found, $F_1(5, 85) = 6.50\ p < .01$ as well as an interaction between the conditions (test with training/retest without training) and type of statement, $F_1(5, 85) = 8.57$, $p < .01$. Since the interaction is due to the under-informative statements with *some,* we limited the analysis to this kind of statement. Children's responses to these statements in the two conditions are significantly different ($t = -3.71, p < .001$). When children were retested a week after the first test and were not given a training session, they accepted the critical statements more ($M = .78$, $SD = .40$) than in Experiment 2 when they benefited from a training session ($M = .466$, $SD = .43$). Of the 10 children that rejected the critical statement more than three times in Experiment 2, only 4 continued to do so in the retest. An analysis by items on the material used in this experiment shows that there is a significant effect of type of statements, $F_2(5, 24) = 45.22)$, $p < .0001$, with the under-informative statements differing from all the others (Scheffé test, $p < .001$).[3]

On average, therefore, the effect of training did not persist, except in a minority of children who were tested. This suggests the following conclusions. First, children's ability to reject under-informative statements is evident only when they are instructed to do so through training. When no training is provided, the majority of children who had previously displayed adult-like behaviour failed to continue to do so, whereas a

[3] In this case, there was also a significant difference between absurd statement and appropriate statements introduced by *some*; between absurd statements and false statements introduced by *all* and between absurd statements introduced by *all* and appropriate statements introduced by *some* (Scheffé test, $p < .05$). In all these cases, the difference was due to the fact that the absurd statements were answered correctly 100% of the time and the other statements were answered correctly between 93 and 95% of the time.

minority of children did hang on to the ability to reject underinformative statements. Although the results from this experiment are therefore weak, they do suggest that the statement evaluation task (evaluating statements without a context) is difficult for children and it is only when specific instructions are given that children take into account the informativeness of the statements. From the results achieved so far, one has to recognise that children's pragmatic competence seems to be hidden in some experiments.

We seem to have reached an impasse. We have seen that children appear to accept under-informative statements. Yet, we have also highlighted the fact that children seem to possess all the necessary prerequisites to derive scalar implicatures, if the criteria for decision-making is reinforced. Moreover, the effect of training is only temporary. One way out of the impasse involves reconsidering the original findings about children's acceptance of under-informative statements. This is the subject of the next section.

EXPERIMENT 4: CONTEXT MATTERS

Let us now consider the factors that contribute to the computation of implicatures, by asking, first, about the contribution of such factors when a person computes an implicature in natural conversation. One factor that is relevant for the computation of implicatures is the awareness that statements can differ in the quantity of information they convey—some statements are more informative than others (e.g., *all*-statements are more informative than *some*-statements). A second factor is the activation of a scale that includes the relevant terms (the use of 'some' activates a scale including 'all'). A third factor is the assumption that a speaker does not utter a *some*-statement when she knows or has evidence that an *all*-statement holds. In turn, the listener assumes that if a *some*-statement is produced by a speaker, then the speaker does not know or has insufficient evidence to conclude that the corresponding *all*-statement is true. If the listener has her own evidence for an *all*-statement, then she can overtly reject the speaker's *some*-statement. In the Noveck study, the third relevant factor for the computation of scalar implicatures was not controlled, in the sense that the evidence for the *some*-statements was not provided in the experimental workspace. In the absence of concrete evidence pertaining to the subjects' judgements, however, we cannot be sure why adults (albeit less than children) sometimes assented to statements like *some giraffes have long necks*. Perhaps the adult subjects took the opportunity to conjure up a subset of giraffes, e.g., baby giraffes, which made the statement express a true proposition. In this case, the evidence against which the statement was evaluated would have included

both baby and adult giraffes; using this domain as the evidential basis, the statement could have been accepted as a reasonable description of the facts. An experimental subject that disagreed with the same statement, by contrast, might have had just prototypical giraffes in mind, i.e., these subjects might have taken the statement to be a description of a typical property of giraffes. In this case the statement would have been an unreasonable description of the facts, since giraffes typically have long necks. Similarly, *some bikes have a handle-bar* might be accepted because the listener might evaluate the statement against the domain including bikes that are broken and lack handle-bars versus bikes that are in good condition. As this discussion shows, the evidence against which statements were evaluated by subjects has not always been controlled in previous experiments; the evidence in question was left up to the subjects to construct. No step was taken to ensure that all subjects evaluated the statements using the same domain, i.e., that of a prototypical individual or entity. Therefore, in the Noveck experiment and in our replication of it (Experiment 1), one possibility is that many children and adults produced a high percentage of acceptance of critical statements because of this feature of experimental design, one that makes the derivation of implicatures irrelevant.

Moreover, on this scenario, it is not surprising that children accepted the critical statements more than adults. First, children may have experienced even more difficulty in figuring out the 'right' or the 'experimenter intended evidence' against which to evaluate the test statements. In addition, children may have been more biased than adults to mentally construct situations that make the experimenter's statements true (for discussion, see Crain & Thornton 1998; Grimshaw & Rosen 1990). To eliminate this potential drawback to the experimental design, we chose to adopt a different methodology in the present experiment, namely one that permits the experimenter to control the evidence that is used by subjects in the evaluation of the test materials. In this case, we can be sure that a potential failure to compute implicatures is not due to the extraneous factors mentioned in the preceding paragraph. The change in method was to opt for the Truth Value Judgement Task (Crain & Thornton, 1998), which is a task that allows the experimenter to control the situation and, thereby, to establish the conditions that are prerequisite for computing scalar implicatures.

Subjects. Fifteen Italian-speaking children (age range 7;0–7;5 years, mean age 7;2) from San Pellegrino Terme (Lombardia) participated in this study and 12 undergraduate students from the University of Milano–Bicocca.

Materials. Subjects were asked to judge five statements including *some* which were true, but under-informative in the context of use (e.g., *Some monkeys are eating a biscuit* in a situation in which all monkeys were eating a biscuit). Since in previous experiments, quantified statements including *some* and *all* in felicitous contexts were not problematic for 7-year-olds, we did not include such statements in the present experiment. To ensure that children could reject false statements and accept true ones, and to be sure that they were paying attention, the targeted statements were interspersed with fillers, some of which were clearly true and others which were clearly false.

Procedure. This experiment used a video-taped version of the Truth Value Judgement Task (TVJT) (Crain & Thornton, 1998). In our version of the TVJT, children watched a video featuring an experimenter acting out stories using props and toys and holding a puppet, Carolina, who was watching the stories alongside the child. At the end of each story, Carolina said what had happened in the story. The child was instructed to say whether Carolina's statement was a good or a bad description of what happened, and to explain her answer whenever she judged Carolina to have 'said the wrong thing.' Children were previously familiarised with Carolina, and had been informed that Carolina was still a baby and, for this reason, she would sometimes be unable to correctly describe what happened in the stories. Children were tested individually in a quiet room where they watched the video together with an experimenter. They were invited to indicate their answer to the experimenter who filled a score sheet and took note of the explanation. Adults were also shown the video and were given a score sheet on which to write their answers. On a typical trial for the target sentence with *some* there were five characters performing some action. For example, one story featured five soldiers that had to go far away to collect a treasure and could either go by motorbike or ride a horse. Initially there was some discussion among the soldiers; some soldiers said that they would like to go by motorbike, since motorbikes are fast; other soldiers argued that gasoline is expensive and that it would be better to ride a horse. After this discussion, they all choose to ride horses. Then, Carolina was asked to say what was happening in the story. In the present case, Carolina's description would be: *Some soldiers are riding a horse*. Then, the child was invited to say whether what Carolina had said was 'right' or 'wrong.'

Results and discussion

The main finding of this experiment is that children rejected the critical statements nearly as often as adults: the rejection rate was 75% for

children and 83% for adults. A one way ANOVA was conducted, with age (Children, Adults) as a factor and logically correct responses as the dependent measure. The analysis revealed that, unlike in Experiment 1, there was no significant difference between children's and adults' responses: $F(1, 25) = 0.31$, $p = .58$ (acceptance of critical statements by children $M = .25$, $SD = .41$; adults $M = .166$, $SD = .389$). On control items, children responded correctly 100% of the time and adults 97% of the time. No reliable difference between these means was found: $F(1, 25) = 2{,}77$, $p = .108$. Finally, the distributions of the responses by children and adults were similar, as shown in Figure 3.

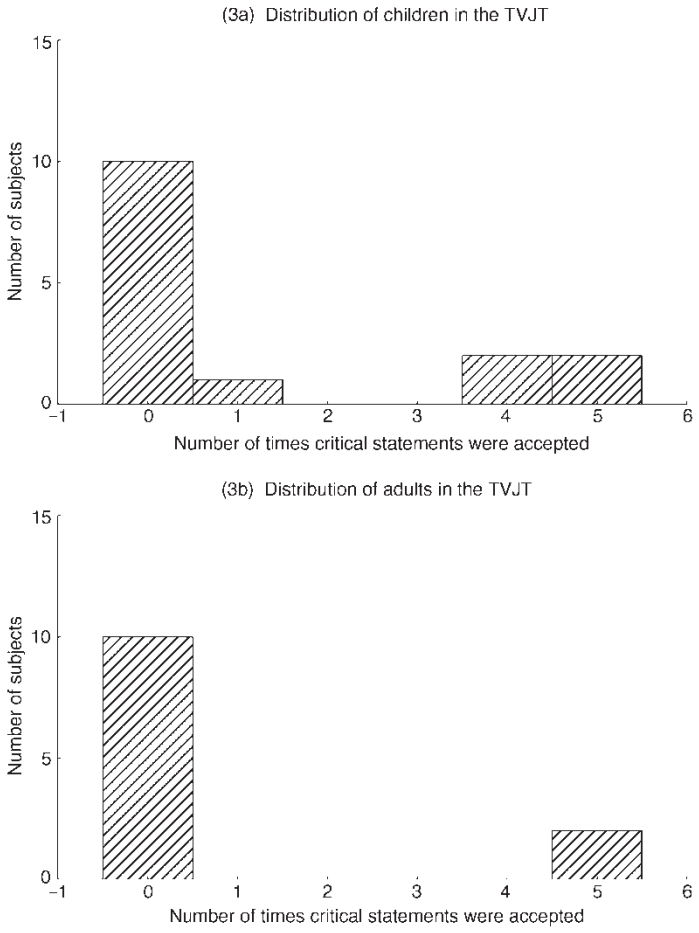

Figure 3. Distribution of children (3a) and adults (3b) depending on the number of times they accepted critical statements in the Truth Value Judgement Task (TVJT).

Most of the subjects always rejected the target statements. The proportions of children and adults who accepted the critical statement three or more times did not differ ($p = .585$, n.s.). Thus, the use of different material and procedure had a dramatic effect; while in Experiment 1 a clear developmental effect was found between children and adults, this effect disappeared completely in this experiment, when a different method was used. Children rejected the critical statements and always explained their responses by invoking the stronger term of the scale; for example they explained that *some soldiers are riding a horse* was a bad description of what was happening because all the soldiers were riding a horse. Thus, we can conclude that 7-year-olds do compute implicatures when the evidence for evaluating under-informative statements is clearly in front of them, and when the task is clear about the nature of judgement that they were required to make. The same observation holds for adults. As in the Papafragou and Musolino (2003) study, our adult subjects rejected the test statements; that is, they computed the implicatures much more often than the corresponding adults had in Noveck's experiment and in our replication of it. In fact, they almost always computed it. We return to this point in the general discussion.

GENERAL DISCUSSION

The current experiments show that 7-year-olds, who are equivocal in judging under-informative factual universal statements (they accept *some giraffes have long necks*), behave like adults and almost always compute the implicatures when the setting is more naturalistic, i.e., when the context makes the relevant evidence immediately available, as in the Truth Value Judgement task (TVJT). This does not appear an *ad hoc* modification of the task; quite the contrary. In ordinary conversational exchanges, speakers and hearers share a common conversational background, which they modify on the basis of what events take place in the context. From the experimental point of view, this means that care has to be taken to ensure that the child and the experimenter also share the same context and conversational background, as happens when one uses the TVJT, but not in the Statement Evaluation Task (SET). In absence of an explicit context and background, children might adopt a different strategy from adults in conjuring up what is intended (see Crain & Steedman, 1985).

All in all, these findings have both theoretical and methodological implications. First, they show that by the age of 7 children consistently derive implicatures, when the context for consideration is made evident to them. In arguably more realistic settings, both children and adults

complain about a speaker's use of a weak scalar term, and they justify their complaints by pointing out that the use of stronger terms would have been more appropriate. Thus, by the age of 7, at least, children's linguistic knowledge of scalar items (*some*) is adult-like. Second, the results suggest that children's failure to derive implicatures in certain experimental conditions does not follow from a lack of pragmatic competence. Rather, the failure may be attributed to the inability to figure out the relevant conversational background for evaluating statements or to a lack of understanding about how the statements are supposed to be evaluated. Third, the fact that adults' responses in the SET (see Experiment 1), but not in the TVJT (Experiment 4), yields a bimodal distribution may indicate that adults perform the task by adopting a strategy which they stick to during the entire experimental session. This, in turn, raises the question of whether children also adopt such a strategy in the relevant experimental conditions.

Let us examine our findings in more detail. First, we extended Noveck's results with French subjects to Italian. As in French, a clear developmental effect was found; children accepted statements like *Some giraffes have long necks* much more frequently than adults did. Then, we showed that children were more ready to reject under-informative statements when the experimental instructions were manipulated using a training session that was intended to make children aware of different degrees of information strength and to enhance their readiness to favour the most informative statement. This manipulation had a dramatic effect on some children, but not on others. Attesting to this is the fact that, in Experiment 2, there was a bimodal distribution of responses across children subjects, when responses were measured as a function of the number of times a child subject accepted the critical statements. Moreover, although training has an effect on a subset of children, the effect did not persist. When children were tested one week later (in a test session that was not accompanied by training), they reverted to accepting under-informative statements much more than they had previously; in fact, about as often as the subjects that were never trained in the first place (Experiment 1). Finally, we tested subjects using a different experimental methodology, the TVJT. The result was a dramatic improvement across all subjects. In this experiment, under-informative statements were rejected almost always, by both children and by adults. In short, the developmental effect observed in the SET (Experiment 1) entirely disappeared. This experimental task showed that, by 7 years of age at least, children can indeed infer implicatures consistently. The TVJT had an stronger effect than the SET plus training (Experiment 3).

It is worth comparing the results of the TVJT and the SET in greater detail. First, in the TVJT, both adults and children rejected under-

informative statements about 80% of the time. In the SET, adults only rejected them around 50% of the time. As we showed, these two means conceal the fact that in the TVJT almost all subjects rejected under-informative statements, thus yielding a unimodal distribution, while in the SET adult subjects were split into two groups, one always rejecting and one always accepting the test statements, yielding a bimodal distribution, a fact that went unnoticed or was not deemed worthy of comment in previous work. But, in our view, the pattern of behavioural responses within and across tasks deserves closer scrutiny. It is important to ask why adults behaved differently in the same task, and why their performance changed depending on the task. There are several related differences between the SET and the TVJT that may answer these questions.

As we remarked above, the TVJT attempts to reproduce an ordinary conversational exchange in which speakers (puppet and experimenter) and hearer (subjects) share a common conversational background (defined by the story), which they update on the basis of what happens in the context (the events occurring during the story). At the end of the story, subjects are asked to say whether the statement used by the puppet to describe what was happening was 'right' or 'wrong.' In short, the TVJT makes all the relevant information for a judgment readily accessible, as is generally the case in normal conversational exchanges. From the responses subjects gave, we infer that almost all adults (and children) evaluated the information strength of the statements with respect to the given context and in so doing they conform to standard conversational norms (see the Introduction): they reject statements with *some* in a situation in which statements with *all* were more appropriate.

By contrast, the set up of SET is not that of an ordinary conversational exchange and does not provide a clear context for the evaluation of the statements, as discussed earlier. Therefore, we are not sure whether adults adopted the standard norms that hold in an ordinary conversation or not. In this situation, adults may have adopted one of two strategies. On one strategy, they might be engaged in an ordinary conversation and adopt accordingly the standard conversational norms. On this strategy, subjects would interpret the experimental instructions to be about the informativeness of the heard description (in an imagined conversational background). Therefore, these adults would be led to interpret *some* as meaning *some, but not all* and disagree with the statements proposed by the experimenter, because they were under-informative. In fact, the adults who rejected the statements explained their disagreement by appealing to the most informative statement. This explains the responses of about half of the adults. The adults that assented to the statements must have adopted a second strategy, with two variants. They might have presumed to be engaged in an experimental task of some sort in which the usual

conversational norms did not hold. If they were to hold, the statement *some giraffes have long necks* would be patently false. They might have reasoned that the experimenter was not asking them to agree or not with a false statement. They might have concluded that either the experimenter had in mind the other meaning of *some* or she was asking subjects to find a context that made the statement not trivially false. In the former case, subjects may conclude that the statements should be taken to mean *some giraffes have long necks and perhaps all* and they would agree with the original statement. Under this view subjects do not compute implicatures because they think they are not at stake, since the norms of conversational exchange do not hold (at least for some subjects). In the latter case, adults attempt to figure out contexts that made these statements sensible (e.g., baby giraffes, broken bikes). Since with some effort and imagination it is possible to find exceptions to factually universal statements and find an appropriate context, some adults might be led to agree with under-informative statements. Notice that given that the context for the evaluation of statements has been enlarged to include marginal cases (e.g., adult and baby giraffes), statements like *some giraffes have long necks* become perfectly informative. Implicatures are inferred, but the statements were true and informative, because of the enlargement of the context. This second strategy could not have been adopted in the TVJT, because the situation was designed to reproduce an ordinary conversational exchange, thus inviting subjects to adopt the conversationl norms; in addition, the context was controlled by the experiment and was made available to subjects (see Experiment 4). Summing up, the adult's response in SET might be explained along the following lines:

(a) Some giraffes have long necks

Adults that disagree	adopt the standard conversational norms and infer the implicature in the usual way and reject (a) as false.
Adults that agree	(i) See that (a) with implicature is false. Conclude that the standard conversational norms, for some reasons (it is an experiment), are not being followed and the 'logical' meaning must be what is intended. (ii) See that (a) with implicature is patently false. Extend the contexts to a non canonical one so as to make (a) true.

Thus, adults in the SET might have adopted a strategy to perform the task and depending on which one they have chosen they were led to always agree or to always disagree with the relevant statements thus giving

responses that yielded a bimodal distribution. These conjectures raise the question of what children do. Sentences are used in conversational contexts, in which speaker and hearer share a background of common knowledge. But the SET does not provide such contexts and children may understand this, but be less able than adults to pretend being engaged in an ordinary conversation. Under this condition, children would be more prone than adults to adopt the second strategy that results in the higher acceptance of underinformative statements. Children either attempt to figure out contexts that make the statement sensible or they assume that the meaning *some and perhaps all* is intended, because they find it awkward that the experimenter reports statements that are trivially false (see Guasti, 2002, chapter 10 for a discussion of infelicities in various experimental set-ups). In any event, the present experiments illustrate very clearly that a certain experimental methodology and certain formulations of task demands may lead to an underestimate of children's pragmatic ability. The change of the methodology, to the TVJT, was effective in making evident children's pragmatic ability, more so even than the combination of training coupled with the SET.

The TVJT has the advantage, as indicated, that it attempts to reproduce a concrete conversational situation, which makes the evidence against which to evaluate statements under the experimenter's control and readily available to the experimental subject. This suggests that crucial factors for enhancing the computation of implicatures are the availability of the relevant evidence and naturalness of the situation. Making subjects aware of the experimental goals (that is, of the fact that they have to judge based on information strength), as we did in Experiment 2 and as Papafragou and Musolino (2003) did, is less of a crucial factor. In this connection, it is worth comparing the target trials with the training items. Recall that in the training session, children frequently offered a specific term like *chef* over a more general term like *man*. This raises the question about the relationship between the training and the experiment. There are two ways to conceive of this relation that rests on different models of implicatures. On one assumption *chef* and *man* form a scale and give rise to implicatures as *all* and *some* do. Thus, the rationale behind the experiment is that training with one kind of scale should improve children's performance with another scale. This is exactly what is found in Papafragou and Musolino's experiment and in our experiment. However, we should be aware that this interpretation of the experimental setting and of the results rests on a particular model of implicatures that is quite controversial. In fact, although *chef* and *man* could give rise to implicatures in specific contexts (see Hirschberg, 1985), it is widely believed that this does not happen in run-of-the-mill contexts, contrary to what happens for *all* and *some*. In particular, in the training session provided in Papafragou and Musolino's

experiment and in our experiment, there does not seem to be anything that would lead one to recognise that by using *man* one intended to deny that *chef* did not apply. This should have happened, however, if we were dealing with a genuine case of scalar implicatures. Notice that when a speaker uses *some*, she is denying that *all* applies (at least as far as she knows). This brings us to the second way of conceiving of the relation between training and test. In line with the neo-Gricean model of implicatures (or any of its more recent developments), one could assume that training simply instructed the children to recognise that *chef* provided a better description than *man* of a given character. Under this view, it remains mysterious why a task in which children are trained to choose the best description should have any effect on the derivation of implicatures.[4,5] It is possible that training alerts some children to be maximally informative.

Our study has shown that some tasks mask children's ability to draw implicatures; when appropriate, 7-year-olds derive implicatures as much as adults do. Thus, the development effect found in Noveck is likely to reflect some difficulty that children experience in evaluating under-informative statements. These difficulties may concern finding out a context for the evaluation of statements or pretend to be in an ordinary conversation, although nothing in the experimental set-up alludes to this. Nevertheless, when we put our study in a larger perspective and compare it to other work in the literature, it becomes evident that implicatures are not regularly computed by younger children than those examined here. Chierchia et al. (2001) tested fifteen 5-year-old American children's interpretation of the scalar term *or* using the TVJT. They found that children assigned to *or* the pragmatic exclusive meaning 50% of the time. By examining the individual subject performance, it surprisingly turned out that the 50% was the result of a bimodal distribution. Children could be divided into two groups: seven children almost always assigned to *or* the exclusive meaning and thus derived the scalar implicature (92%), while another group of seven children almost never did (7%). Only one child behaved at chance. While it was clear that age was not the critical factor that distinguishes the two

[4] A less controversial training would have been one in which the terms used formed a scale under any model of implicatures (e.g., the connectives *and*/*or*).

[5] Based on these considerations, one might raise doubts on the view that training enhances subjects' readiness to derive implicatures. It is possible that training merely serves to instruct subjects to search for statements that describe the observable facts more perspicuously, without necessarily leading them to really derive implicatures, that is, to recognise that by uttering *some giraffes have a long neck* the speaker intended to communicate that *some, but not all giraffes have a long neck*. Although it is hard to distinguish between the search of a more adequate description and the derivation of implicatures, these concerns should not be dismissed.

groups of children, it remains mysterious why 5-year-olds behaved differently. Papafragou and Musolino (2003), also using the TVJT, found that 5-year-old Greek-speaking children did not regularly compute implicatures.[6] This brings us back to Noveck's finding. Although 7-year-olds do not differ from adults in the ability to draw implicatures, it is possible that there is a developmental effect when we consider 5- or 6-year-olds. At this point, our hypotheses can only be speculative and further testing is needed. It is possible that 5- to 6-year-old children are not able to derive pragmatic inferences, because the weaker statement (with *some*) does not elicit the activation of the alternative strongest statement (*all*) or that scalar items are not in a scale, a piece of knowledge that may be acquired as part of the acquisition of the lexicon. In this case, children do not see that 'some XP are YP' is under-informative, because they fail to recognise that the utterance of *some* implicates *not all*. Alternatively, it is possible that the strongest statement is activated, but the implicit reasoning that leads one to recognise that *some* implicates *not all* is not carried out by children, possibly because this would exceed children computational resources. Further experiments are needed to investigate these hypotheses. The aim of our study was to show that to understand how pragmatic abilities develop, we must be sure that children (and adults) and the experimenter share the same conversational context. This might have been a source of confounding in previous experiments. When the conversational context was controlled for, 7-year-olds, who were equivocal in deriving implicatures, were not hesitant in rejecting *some monkeys are eating a biscuit* in a situation in which all the relevant monkeys were eating a biscuit. Of course controlling the conversation background is one factor that may be relevant for deriving implicatures or not. It is still possible that differences among adults and children, and among children of different ages remain even after this manipulation. In this case, we can feel confident in attempting to ascribe the differences to the various components entering in the derivation of implicatures (activation of strongest statements, recognition that *some* implicates *not all*).

Manuscript received May 2003
Revised manuscript received July 2004

[6] Unfortunately, these authors only provide means and result from the ANOVA and do not tell us how subjects are distributed depending on the number of times they accepted under-informative statements. But we guess that subjects' responses did not have a normal distribution.

REFERENCES

Braine, M. and Roumain B. (1981). Children's comprehension of ''or'': Evidence for a sequence of competencies. *Journal of Experimental Child Psychology*, *31*, 46–70.

Chierchia, G. (2004). Scalar implicatures, polarity phenomena, and the syntax/pragmatic interface. In A. Belletti & L. Rizzi (Eds.), *Structures and beyond*. Oxford: Oxford University Press.

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. Dominguez, & A. Johansen *Proceedings of the 25th Boston University Conference on Language Development.* Sommerville, MA: Cascadilla Press.

Crain, S. & Steedman, M. (1985). On not being led up the garden path: The use of context by the psychological syntax processor. In D. R. Dowty, & L. Karttunen, & A. Zwicky (Eds.) *Natural language parsing: Psychological, computational and theoretical perspective*. Cambridge: Cambridge University Press.

Crain, S. & Thornton, R. (1998). *Investigations in universal grammar. A guide to experiments on the acquisition of syntax and semantics.* Cambridge, MA: MIT Press.

Grice, P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.) *Syntax and semantics 3: Speech acts*. New York: Academic Press. Also in Paul Grice (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.

Grimshaw, J. & Rosen, S. T. (1990). Knowledge and obedience: The developmental status of the binding theory. *Linguistic Inquiry*, *21*, 187–222.

Guasti, M. T. (2002). *Language acquisition. The growth of grammar*. Cambridge, MA: MIT Press.

Horn, L. (1972). *On the semantic properties of logical operators in English*. Doctoral Dissertation, UCLA, CA.

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

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

Papafragou A. & Musolino, J. (2003). Scalar implicatures at the semantic-pragmatics interface. *Cognition*, *80*, 253–282.

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

APPENDIX A

Material used in Experiment 1

Bizarre	Factually universal	Factually existential
All birds have telephones.	All hammers have a handle.	All dogs are black.
All crayons have noses.	All books have pages.	All animals leave in the water.
All chairs play instruments.	All pigeons have wings.	All pants are short.
All doors sing.	All elephants have trunks.	All birds live in cage.
All couches have windows.	All refrigerators have doors.	All cars are red.
Some fishes are made of leaves.	*Some bikes have wheels.*	Some flowers are yellow.
Some oranges have computers.	*Some cars have motors.*	Some dresses have pockets.
Some books are good to eat.	*Some giraffes have long necks.*	Some chairs are made from wood.
Some stores are made out of babbles.	*Some cats have ears.*	Some children are blond.
Some children are made out of feathers.	*Some airplanes have wings.*	Some cakes are made from chocolate.

Material used in Experiment 3

Bizarre	Factually universal	Factually existential
All tables have ears.	All pots have handles.	All skirts are short.
All windows talk.	All horses have tails.	All dogs are black and white.
All mice have cars.	All books are made out of paper.	All shoes are made out of canvas.
All beds have doors.	All dauphins live in the water.	All children have blue eyes.
All wardrobes laugh.	All balls are round.	All apples are yellow.
Some dogs speak French.	*Some cats have hair.*	Some women have blond hair.
Some children are made out of bricks.	*Some cows have eyes.*	Some shirts are red.
Some pears bike.	*Some houses have a roof.*	Some tulips are yellow.
Some houses are made with sugar.	*Some bikes have handle-bars.*	Some dishes are made out of plastic.
Some glasses are good to drink.	*Some buses have wheels.*	Some cakes are covered with marmalade.

Material used in Experiment 4

Critical statement	Fillers
Some monkeys are eating a biscuit.	Mommy bear is buying some bananas.
Some soldiers are riding a horse.	Batman is bringing some flowers.
Some girls are watching TV.	The lady is taking a bath.
Some dwarfs are going on a boat.	Ninja is sleeping.
Some clowns are fishing.	