Ad-hoc Implicature in Preschool Children

Alex J. Stiller

Department of Linguistics, University of California, San Diego

Noah D. Goodman

Department of Psychology, Stanford University

Michael C. Frank

Department of Psychology, Stanford University

Thanks to Janelle Klaas for assistance in data collection and to Cindy Fisher and three anonymous reviewers for helpful suggestions. A previous version of this work was presented to the Cognitive Science Society in Stiller, Goodman, and Frank (2011).

Please address all correspondence to Michael C. Frank, Department of Psychology, Jordan Hall (Bldg. 420), 450 Serra Mall, Stanford, CA 94305. Phone: (650) 724-4003. E-mail: `mcfrank@stanford.edu`

Abstract

If a speaker tells us that "some guests were late to the party," we typically infer that not all were. Implicatures, in which an ambiguous statement ("some and possibly all") is strengthened pragmatically (to "some and not all"), are a paradigm case of pragmatic reasoning. Inferences of this sort are difficult for young children, but recent work suggests that this mismatch may stem from issues in understanding the relationship between lexical items like "some" and "all," rather than broader pragmatic deficits. We tested children's ability to make non-quantificational pragmatic inferences by constructing contextually-derived "ad-hoc" implicatures, using sets of pictures with contrasting features. We found that four-year-olds and some three-year-olds were able to make implicatures successfully using these displays. Hence, apparent failures in scalar implicature are likely due to difficulties specific to the constructions and tasks used in previous work; these difficulties may have masked aspects of children's underlying pragmatic competence.

Introduction

Human communicators typically follow general principles of cooperation, such as being truthful, relevant, perspicuous, and adequately informative. By assuming that their partners abide by these conventions, listeners can draw inferences about the meanings speakers intend to convey (Grice, 1989; Hirschberg, 1991; H. Clark, 1996; Horn, 1998; Levinson, 2000). For example, consider the following exchange:

(1) A: Did you visit your cousins?
B: I visited some of them.

In (1), A may infer that B did not visit all of her cousins. If she had visited all of them, the word "all" would have been the maximally informative choice. Even though saying "some" would have been true even if she had visited all of them, the choice of "some" suggests that she has chosen not to say "all," likely because it is not true. B's inference, that A's intended meaning ("some but not all") is more restricted than the literal meaning of her utterance ("some"), is an example of a *pragmatic implicature.*

Grice (1975, 1989) introduced a distinction between two types of implicatures: *generalized* and *particularized*. Generalized implicatures, also commonly known as *scalar implicatures* (or SIs, the label we adopt throughout the manuscript), involve lexical items that are ordered with respect to one another, including but not limited to quantifiers (<SOME, ALL>), modals (<MIGHT, MUST>), and numerals (<ONE, TWO>). A detailed description of the ordering relations among such terms is given by Horn (1998).

In contrast to SIs, *particularized*, or *ad-hoc*, implicatures are cases in which an inference is available due to special features of the context. The important distinction between the two types of implicatures is that in ad-hoc cases, the relationship between alternatives relies on context whereas in generalized cases, the set of alternatives is a feature of the language more generally. While some theories emphasize the differences in computation between these types of implicature (e.g. Levinson, 2000), others minimize them (e.g. Sperber & Wilson, 1986). We remain agnostic about the issue; on all accounts, generalized implicatures differ from particularized implicatures minimally in that they require knowledge of the lexical alternatives (e.g. "some," "all") that constitute the scale (though there may be other relevant differences).

Our goal here was to measure preschool children's ability to make ad-hoc (particularized) implicatures. Even older children have been reported to have difficulty with scalar implicatures. Thus, measuring children's performance in ad-hoc cases can contribute to an understanding of whether failures with scalar implicature are due to specifics of these implicatures per se. If children succeed in making ad-hoc implicatures at a younger age than they perform scalar implicatures, this evidence would rule out broader pragmatic deficits, such as difficulties computing informativity or going beyond what is said. To ground this discussion in previous research, we review the literature on implicature in development below before describing the specifics of our experiment.

Implicature in Development

Implicatures—especially SIs like (1)—have been taken as a paradigm case of pragmatic inference, and their development has been a subject of considerable interest (Braine & Rumain, 1981; Papafragou & Musolino, 2003; Huang & Snedeker, 2009; Barner, Brooks, & Bale, 2011). A number of experiments have suggested that SIs, especially those involving the quantifiers "some" and "all," are difficult for children until late in development. In one influential study, Noveck (2001) reported difficulties involving modal operators such as "might" and "must." These paradigms were both relatively complex, however, requiring not just an understanding of implicature, but also an understanding that implicatures could render a statement infelicitous (e.g. "some dogs are animals") and that such a statement should be judged false. Even if children made the scalar implicature in such a case, they might not have taken the step of assuming that the possible implicature necessarily made the original statement false.

Stronger evidence comes from a series of foundational studies on SI interpretation (Papafragou & Musolino, 2003; Huang & Snedeker, 2009). In one of these, Huang and Snedeker (2009) measured eye movements of children and adults as they listened to SIs. Participants saw various scenarios corresponding to weak and strong interpretations of scalar terms, and their relative looking time to the scenes was measured after they heard a reference such as "the girl who has some of the socks." While adults eventually generated the SI inference (albeit after a delay; though cf. Grodner, Klein, Carbary, & Tanenhaus, 2010), five-year-olds did not. Furthermore, adults, but not children, were able to distinguish between scenarios that were consistent with an implicature (e.g. when "the girl that has some of the socks" described a character with two of the four socks in the display) and those that violated it (e.g. when "the girl that has some of the socks" described a character with all four of the four socks in the display). These findings provide the clearest evidence to date that SIs with quantifiers are difficult for children.

These findings are surprising with respect to the broader developmental literature for at least three reasons: First, there is a large and consistent body of evidence that children learn new words by relying on their understanding of the goals and intentions of others, i.e. they learn words "pragmatically" (Baldwin, 1993; Tomasello & Akthar, 1995; Bloom, 2002; Frank, Goodman, & Tenenbaum, 2009; E. V. Clark & Amaral, 2010). If children do in fact use pragmatic reasoning to learn new words, why can't they use that knowledge to compute SI inferences? Second, an increasingly broad literature suggests that toddlers and infants can reason about both the goals (Gergely, Bekkering, & Király, 2002; Meltzoff, 1995; Woodward, 1998; Gergely et al., 2002) and beliefs (Onishi & Baillargeon, 2005; Southgate, Senju, & Csibra, 2007; Buttelmann, Carpenter, & Tomasello, 2009) of other agents. Third, some of the precise abilities involved in pragmatic reasoning—in particular a sensitivity to informativeness—are present in younger children as well. While these first two reasons are speculative, relying on potential links between implicature on the one hand and social cognition and word learning on the other, the last is more directly relevant and bears more detailed explanation.

A variety of evidence suggests early sensitivity to informativeness on the part of children (e.g. see Chierchia, Crain, Guasti, Gualmini, & Meroni, 2001; Foppolo, Guasti, & Chierchia, 2012). At age three, children are more likely to produce informative referring expressions when interlocutors are blind to a scene (Matthews, Lieven, Theakston, & Tomasello, 2006), and at four, they are more likely to provide more information in descriptions when distractors are similar to a target (Matthews, Butcher, Lieven, & Tomasello, 2012). By age five, when they are still failing many scalar implicature tasks, children show sensitivity to the informativeness of speakers' statements in the rewards they give (Katsos & Bishop, 2011) and include supplementary adjectives when needed to identify a target referent unambiguously (Nadig & Sedivy, 2002).

Results with pointing gestures are even stronger. Twelve-month-olds point to identify the location of a target object unambiguously (Liszkowski, Carpenter, Striano, & Tomasello, 2006), and two-year-olds know when their own pointing gestures do not uniquely identify a referent and adjust their communication strategies accordingly (O'Neill & Topolevec, 2001; see also Liszkowski, Carpenter, & Tomasello, 2008, Matthews, Lieven, & Tomasello, 2007, Matthews et al., 2012). Taken together, these findings suggest an early understanding of informativeness even in production, which typically lags behind comprehension. If SI follows from an understanding of informativeness (Horn, 1998; Hirschberg, 1991; Levinson, 2000), then children who know what is—and is not—adequately informative should be able to use that knowledge to compute SI inferences, in the absence of other obstacles.

This reasoning has driven a number of authors to consider other factors that might cause children's failure in SI tasks (Noveck, 2001; Huang & Snedeker, 2009; Barner & Bachrach, 2010). These include difficulties accessing relevant the lexical alternatives (e.g. "all" when "some" is mentioned; Chierchia et al., 2001; Barner & Bachrach, 2010), and knowing that one alternative in SI tasks negates others (Barner et al., 2011). Apparent failures may also be due to the methodologies of truth-value or felicity judgment (Guasti et al., 2005; Papafragou, 2006). These methods, which ask children to judge whether an implicature violation is felicitous or correct, cannot differentiate failure to compute SIs from general tolerance of pragmatic violations (Katsos & Bishop, 2011).

Barner et al. (2011) conducted an experiment that tested whether access to lexical alternatives posed a problem for children in computing SIs. They showed children displays where a property was true of some or all of the members of a set, for example a group of three animals in which all three were reading. In the critical conditions, the majority of 4 – 5 year-olds endorsed the pragmatically infelicitous "some" in a context where "all" could have been used, consistent with previous work on SI. But they also endorsed the logically false statement that "only some" were sleeping. In contrast, when the animals were enumerated (e.g. "only the cat and the cow are sleeping"), children correctly rejected this statement in cases where the modifier "only" made it false, suggesting that they understood what "only" meant. Barner and colleagues interpreted this set of results as suggesting that children were unable to call to mind "all" as the scalar alternative to "some," even when it was grammatically required by the word "only." This interpretation provides a plausible explanation for previous failures: although children may have been able to understand that "some" was not maximally informative, they nevertheless could not summon the relevant alternative to mind to compute a SI.

While Barner et al.'s (2011) study explains children's failures, there are as yet only limited positive demonstrations of any implicature abilities in children younger than 5, even though such demonstrations should in principle be possible. Miller, Schmitt, Chang, and Munn (2005) asked children to select a picture in which a puppet made "some faces happy" by drawing smiling mouths on some but not all of the available faces (distractor items included an "all" picture and a "none" picture). In a condition when "some" was stressed, children chose the SI-consistent "some" picture but not the "all" picture, while they chose both pictures together most often in the unstressed condition. This result suggests that a referent-selection task might be promising for eliciting successful implicatures, but the small sample in each condition (N=8) and broad age range (3;6 – 5;10) limit the strength of the inferences that can be made from this study.

Papafragou and Tantalou (2004) also provided some evidence that children could compute implicatures (both quantificational and ad-hoc), in this case in a competitive felicity judgment task. Children saw e.g. a tiger who was assigned to eat a set of oranges, and who reported "I ate some" (in the quantifier condition) or a cow who was assigned to wrap a set of gifts (a parrot and a doll) and reported "I wrapped the parrot" (in the ad-hoc condition). While a group of 10 children (mean age 5;2) correctly awarded or withheld prizes from the puppet based on the performance implied by these statements, these children were on average fully 20 months older than the 3-year-olds we consider here. More importantly, new evidence from Sullivan, Davidson, and Barner (2011) suggests that the children in Papafragou and Tantalou (2004) may have succeeded purely by relying on the Principle of Contrast—giving prizes when the reported action exactly matched the assigned action, and failing to give prizes when the reported action contrasted—rather than computing any pragmatic implicature (E. V. Clark, 1988); thus, Papafragou and Tantalou (2004)'s results should be interpreted with caution.

The Current Study

In sum, previous work has suggested that SI inferences—and perhaps pragmatic inferences more generally, though the evidence on this issue is more limited—are difficult

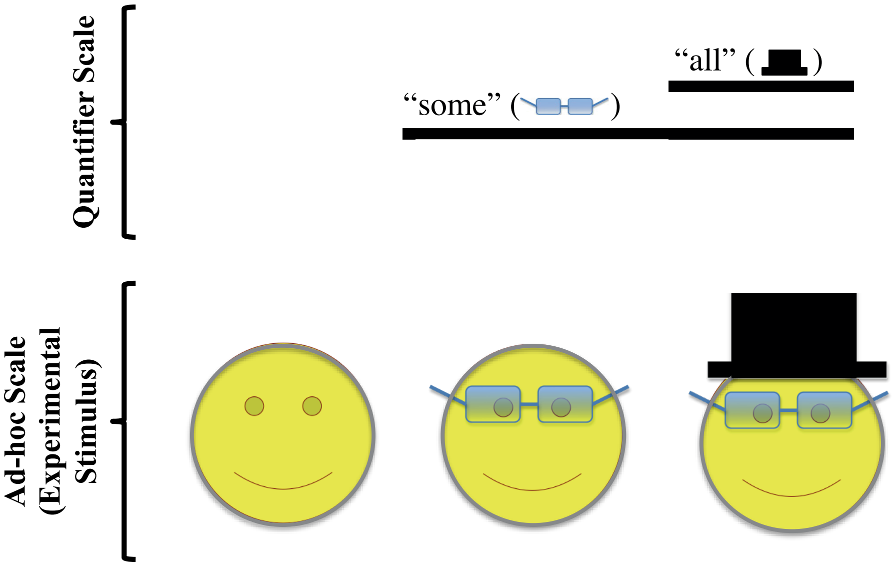

Figure 1. Example stimuli from our referent selection task. The middle item represents the pragmatically enriched interpretation of "My friend has glasses." The implicature has a similar logical structure to the conventional *some-not-all* implicature (top).

for young children. But in light of the arguments reviewed above, such findings present at best an ambiguous picture of children's pragmatic reasoning abilities. A positive demonstration of ad-hoc implicature in younger children would help to clarify this picture by suggesting that the challenges of scalar implicature do not extend to this domain. The current work attempts to provide such a demonstration.

We created a novel, child-friendly referent-selection paradigm, pictured in Figure 1. In this context, a speaker who asserts, "My friend has glasses" (in the experimental, "label" condition) implies that her friend is wearing only glasses. This inference is formally identical to the quantificational inference that "some" implies "only some," but to facilitate comprehension, our stimuli create an ad-hoc scale in which scalar alternatives are concrete nouns ("hat" and "glasses") rather than abstract lexical items such as quantifiers or modal operators. Our paradigm also mitigates possible difficulties associated with calling to mind relevant alternatives by presenting the possible referents side by side. Finally, to address the fact that truth-value judgment tasks and felicity judgment tasks do not distinguish between pragmatic failures and mere pragmatic tolerance (Katsos & Bishop, 2011), our referent-selection task conveys that when one alternative is selected, the others cannot be the case (Barner et al., 2011).

As a control against baseline differences, we created a further "no label" condition in which we asked children to choose one stimulus (e.g. the "friend") but did not give any further information. This method, referred to in Frank and Goodman (2012) as the "contextual salience" method, allows us to measure children's baseline belief that one or the other of the items is most likely to be the puppet's intended referent. If children in the experimental condition are making pragmatic enrichments to their linguistic input, then we would expect them to pick the target (i.e. single-feature) item less frequently in the absence of this input.

Methods

In this report, we provide data from two independent samples. We initially collected a planned sample of 24 children per age group across three ages (2 – 3 year olds, 3 – 4 year olds, and 4 – 5 year olds) and two conditions ($N_{sample\ 1} = 147$). Due to the loss of video tapes and records for a subsection of the sample we were unable to recode participants' responses for their choice patterns (see below). We therefore conducted an independent replication with a second planned sample ($N_{sample\ 2} = 144$).

Participants

Data in the first sample were collected from 147 children: in the label condition, 25 two-year-olds (M=2.6 years), 26 three-year-olds (M=3.5 years), and 24 four-year-olds (M

= 4.5 years) participated at Bing Nursery School of Stanford, CA and the Children's Discovery Museum (CDM) of San Jose, CA. In the no-label condition, an additional 24 two-year-olds (M=2.6 years), 24 three-year-olds (M=3.5 years), and 24 four-year-olds (M=4.5 years) participated at the same locations. Data in the second sample were collected from 144 children, all recruited at the CDM. In the label condition, there were 23 two-year-olds (M=2.6 years), 24 three-year-olds (M=3.5 years), and 25 four-year-olds (M = 4.5 years), and in the no label condition there were 24 two-year-olds (M=2.5 years), 24 three-year-olds (M=3.5 years), and 24 four-year-olds (M = 4.5 years). Experimenters recruited children for a "storybook activity." Parents were present during data collection at CDM, and they watched quietly from across the room.

In the second sample, 33 additional children contributed data but were not included in the final sample because of reported English exposure in the home being less than 75% (25), because of parental interference (5), because they failed to complete the study (2), or because they had a self-reported developmental language disorder (1).

Adults in the label and no-label conditions were 48 participants recruited using Amazon's Mechanical Turk web-based crowd-sourcing platform (24 in each condition).

Stimuli

Stimuli for children were arranged in a binder containing materials for six trials: four inference trials (such as those described above and pictured in Figure 1) and two filler trials.

The unambiguous filler trials consisted of three different colored cars and three different kinds of fruit. These trials were included as a check to ensure comprehension. In the second sample, three children each made a single mistake on a filler trial (one two-year-old, one three-year-old, and one four-year-old), yielding 98% performance overall. We do not discuss the filler trials further.

In each inference trial, three copies of the same base object were present, with two features varying across the set. Inference trial materials were sets of faces (with glasses and hats as features), houses (with trees and flowers), plates of pasta (with meatballs and sauce), and beds (with a teddy bear and a stuffed penguin). One object from the base set had neither feature ("distractor"), one had exactly one feature ("one-feature"), and one had both features ("two-feature").

Positions of the three objects and which feature was used for the one-feature object (e.g. only hat vs. only glasses) were counterbalanced such that the position of each item and named feature occurred an identical number of times. To accomplish this, six orders were necessary. The assignment to one of the six orders was random and identical for the children in the control and test conditions. All orders began with a filler trial to ensure that children understood the task.

Procedures

The task was administered by an experimenter, who used a stuffed animal as a confederate. In the first sample, the stuffed animal was a green monster named "Furble," while in the second sample, the stuffed animal was a red dog named "Clifford." The experimenter asked participants to help the stuffed animal identify various people and objects. In the Label condition (experimental), each inference trial consisted of the stuffed animal using a description, "My X is/has Y," that was ambiguous between the one-feature and two-feature objects. For example, in a trial like the one pictured in Figure 1, the stuffed animal would say "My friend has glasses." Children were then asked to point to the appropriate item and their response was recorded. On filler trials, the stuffed animal simply referred to one of the items unambiguously, e.g. "My car is red."

In the No Label (control) condition, the procedure was identical, but children heard a revised story in which the stuffed animal would say something unintelligible. In the first sample, the cover story was that Furble had eaten too much peanut butter to speak. In the second sample, the cover story was that Clifford was a dog and could only bark. Thus, instead of saying, “My friend has glasses,” the stuffed animal would simply mumble or bark. Children were again asked to pick out the item they thought belonged to the stuffed animal. There was no correct answer on filler trials in the no-label condition.

Adults completed an equivalent task embedded in a webpage, picking alternatives from each set of objects by clicking on corresponding radio buttons. The adult version used the same script with a picture of Furble substituting for the stuffed animal in the live action version. Adults in the control condition saw strings of hash marks instead of the names of features. Adult participants were informed that the task was designed for children.

Results

The primary question of interest in our analysis was whether participants’ choices indicate a successful pragmatic inference: in other words, whether they chose the one-feature object (e.g., the face with glasses but no hat) in contrast to the two-feature object (e.g., the face with glasses and a hat). We begin our analysis by examining the influence of different factors on this primary measure of interest (one-feature choice); subsequently we consider different ways of answering the question of whether participant judgments reflect pragmatic inference. We end by considering our adult control data.[1]

Initial Analyses

Because of the large number of participants we tested, we were able to divide our sample into half-year age groups. Means and standard deviations for these age groups are given in Table 1 and are plotted in Figure 2. Throughout this section we used logistic

[1] Data and code used in these analyses are available at `http://github.com/langcog/scales`.

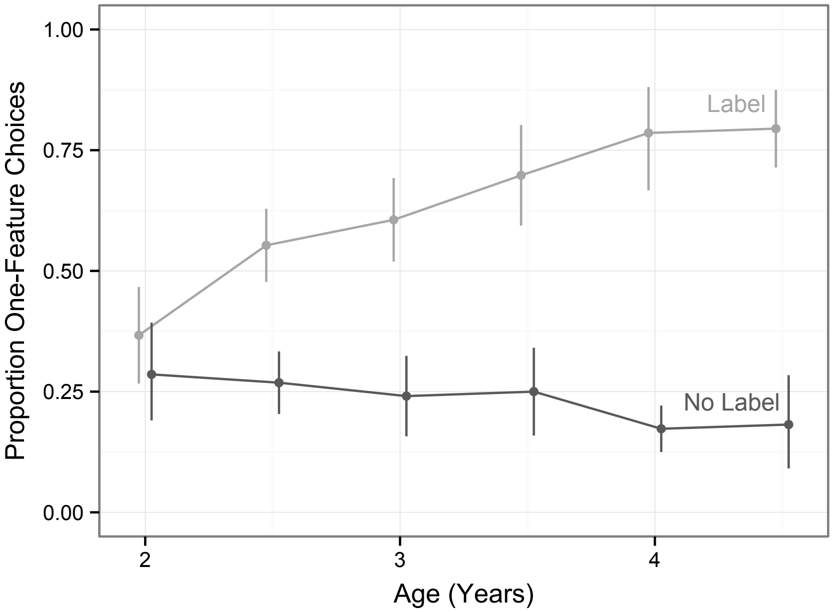

Figure 2. Mean proportion of choices indicating the one-feature (implicature-consistent) object on inference trials in the Label (light gray) and No Label conditions (dark gray) across ages. Error bars show 95% confidence intervals computed via subject-wise non-parametric bootstrap.

mixed-effects models to quantify effects of different factors on implicature-consistent (one-feature) responding. In all of these models, we estimate the influence of various factors on this dependent variable with crossed random effects of participant and item (Gelman & Hill, 2007), the maximal random effects structure justified by our experimental design (Barr, Levy, Scheepers, & Tily, 2013).

Our first analysis used a model that included age (as a continuous factor), condition, and their interaction. We found a positive effect of age ($\beta = .72$, $p < .0001$), a negative effect of the No Label condition ($\beta = 1.65$, $p = .005$), and a negative interaction of the two ($\beta = -.99$, $p < .0001$). In other words, older participants were more likely to choose the implicature consistent response, but primarily in the Label condition. An

Label (Experimental)				No Label (Control)			
Age	Mean	SD	N	Age	Mean	SD	N
2.0	0.37	0.21	15	2.0	0.29	0.23	21
2.5	0.55	0.22	33	2.5	0.27	0.20	27
3.0	0.61	0.24	26	3.0	0.24	0.23	27
3.5	0.70	0.28	24	3.5	0.25	0.23	22
4.0	0.79	0.27	21	4.0	0.17	0.14	26
4.5	0.79	0.22	28	4.5	0.18	0.23	22

Table 1

Summary statistics for one-feature responses by age group for each condition. Age indicates half-year age bins (e.g. 2.0 indicates children from 2 years 0 months to 2 years 6 months).

examination of the random effects suggested that there was some item-level variation ($\beta_{beds} = .17$, $\beta_{faces} = -.04$, $\beta_{houses} = -.16$, $\beta_{pasta} = .04$), but all four items showed the same basic developmental trends.

We examined differences between our two independent samples by adding sample as a factor to our previous model, and adding all two- and three-way interactions between sample and other variables. We found that there was a negative coefficient for the second sample that was almost reliable ($\beta = -1.49$, $p = .06$) but no reliable interactions with sample ($p > .14$). This trend towards a main effect of sample suggested slightly lower performance in choosing the implicature-consistent target for children in the second sample. We speculate that this may be due to the composition of the first sample, which included some children from an on-campus nursery school where younger children especially may have felt more comfortable in the testing situation. Nevertheless, a model

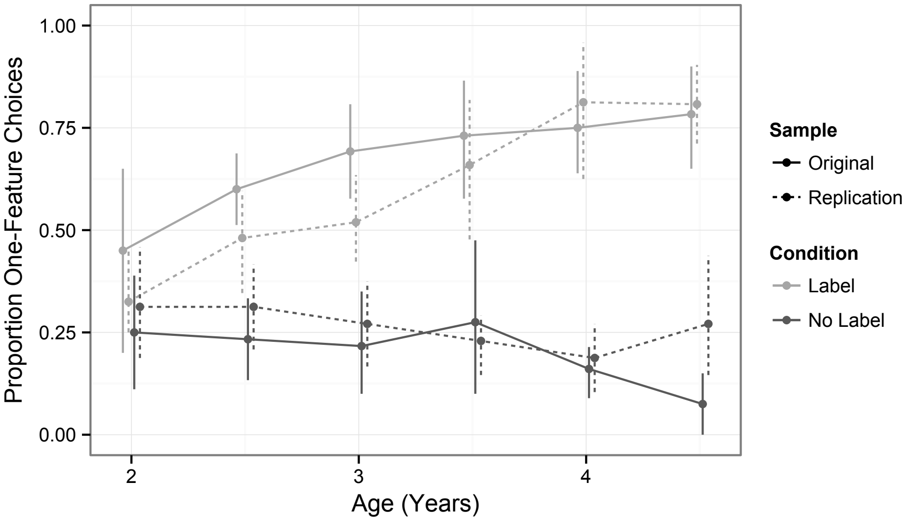

Figure 3. Mean proportion of choices indicating the one-feature (implicature-consistent) object on inference trials in the Label (light gray) and No Label conditions (dark gray) across ages, divided by sample. Error bars show 95% confidence intervals computed via subject-wise non-parametric bootstrap.

that included only the second sample showed exactly the same pattern as the model that included both, with reliable effects of age ($\beta = .87$, $p < .0001$), condition ($\beta = 2.33$, $p = .003$), and their interaction ($\beta = -1.09$, $p < .0001$). The two samples are compared in Figure 3.

We next modeled the effects of demographic factors on responding in our second sample, for which we had available a short demographic information sheet given to all parents participating in research at Children's Discovery Museum. A model including gender showed a reliable negative coefficient for males' responses ($\beta = -.94$, $p = .02$) suggesting that male children made fewer implicature-correct responses; there were no reliable interactions between gender and age or condition. We did not find an effect of self-reported percentage exposure to English in the home (main effect $\beta = .003$, $p = .80$,

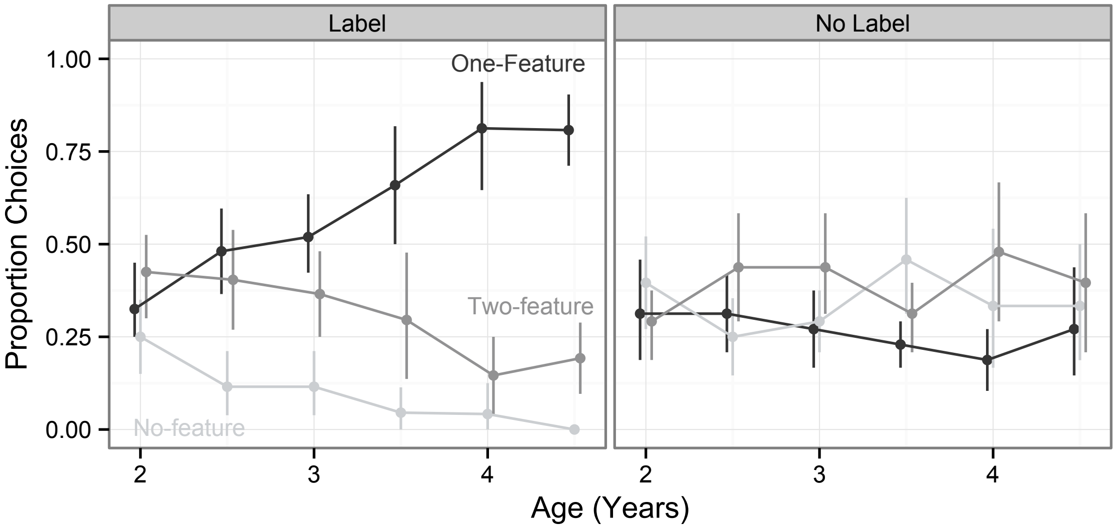

Figure 4. Mean proportion of choices indicating the one-feature (implicature-consistent), two-feature, or no-feature object on trials in the Label and No Label conditions, across ages. Data are from second sample only. Error bars show 95% confidence intervals computed via subject-wise non-parametric bootstrap.

no reliable interactions), presumably because we excluded children with percentages lower than 75% and hence our range was restricted. We also did not find any main effect or interaction with parent education (main effect $\beta = -0.03$, $p = .69$).

To summarize, we found a consistent pattern in our data: a developmental increase in responding in the Label condition but not the No Label condition. The developmental increase was marginally earlier in our first sample than our second, and appeared to be stronger for girls than for boys, but overall all analyses yielded a consistent picture of the data.

Evidence for Pragmatic Inference

We next turn to an examination of whether—and when, developmentally—our data yield evidence of pragmatic inference. In order to do so, we start with the observation that there are two differences between the Label and No Label conditions. First, in the Label condition, children can use the name of a feature (e.g. "glasses") to narrow down the reference set logically, that is to the two objects that possess that feature. Second, they can make a pragmatic inference that the name refers to the object with *only* that feature. It is only this second difference that we are interested in.

To examine this second factor, we look to the distribution of children's responses across possible targets (Figure 4). (All analyses in this section use only the second sample, for which response data was available). In the Label condition, the youngest 2-year-olds were close to random in their responses. For the 2.5 – 3 year-olds, responses to the no-feature object were lower and responding was approximately even between the logically-possible alternatives. In the 3 – 3.5 year-olds, responding was noticeably higher for the one-feature than the two-feature object (52% vs. 37%); and this pattern was even stronger for older groups.

We next attempt to quantify these differences. In these analyses, we model only the subset of responses in which children did not choose the no-feature distractor, again using a logistic mixed-effects model. Chance responding for this analysis was 50%, and a reliable positive coefficient is a signal of pragmatic inference. We pursued two alternative modeling approaches: modeling the Label condition independently, and modeling the two conditions jointly.

Pursuing the first approach, we used a mixed model to compare one-feature responding to chance in the Label condition alone, reasoning that individual participants could show clear evidence of greater-than-chance responding by choosing the one-feature response at above chance levels. In this analysis, the youngest age group for which there was a significant bias to choose the one-feature object was the 3.5-year-olds ($\beta_{3.5} = 1.07$, $p = .03$); 4 and 4.5 year olds were highly reliable in their one-feature responding ($\beta_{4.0} = 1.99$, $p = .0003$ and $\beta_{4.5} = 1.70$, $p = .0008$ respectively).

On the other hand, the reason we included the No Label condition in our experiment was to provide a baseline measurement of the salience of different alternatives (Frank & Goodman, 2012). In previous work, we have shown that this salience can affect participants' baseline responding. Thus, a more sensitive test for pragmatic inference might be having overcome that baseline responding bias. To quantify this effect, we fit a logistic mixed model with data from both the Label and No Label conditions (again, excluding no-feature responses). This model contained both coefficients for each age group and interactions between age and condition. We set responses in the No Label condition as the baseline; thus an interaction between condition and age group would be a signal of greater-than-baseline responding for a particular condition. In this model we found a trend towards such an interaction in the 3.0 age group ($\beta_{3.0} = .83$, $p = .07$), indicating that there was some baseline bias that participants might be overcoming. Coefficients for all subsequent age groups were reliable ($\beta_{3.5} = 1.11$, $p = .03$; $\beta_{4.0} = 2.66$, $p < .0001$; $\beta_{4.5} = 1.81$, $p = .0003$).

In summary, we found reliable evidence for pragmatic inference beyond the literal interpretation of a linguistic description in children from 3.5 – 4 years old, with suggestive evidence of an effect in 3.0 – 3.5 year-olds. In contrast, we saw no such evidence in two-year-olds. Was this lack of positive responding due to difficulties that the two-year-olds had with the task? The number of trials might have been taxing on the attention spans of two-year-olds, and we did not actively control the level of the vocabulary items that were used in the stimuli. In addition, we saw substantial (25%) incorrect responses (choosing the no-feature item) for the youngest two-year-olds in the Label condition, suggesting that they were not able to succeed in the basic language interpretation component of the task with high reliability.

In a pilot follow-up experiment, we created a version of the task that tested the contribution of these factors to younger children's performance. In this version, we included one fewer trial, added a common grounding phase in which objects were named (Barner et al., 2011; Papafragou & Musolino, 2003), and used only vocabulary items that were very likely to be known to young children. We still saw no sign of above-chance responding from a group of 12 two-year-olds, so we suspect that for younger children there may be obstacles to success in the particular paradigm we used here. One such obstacle is that the one-referent implicature target is relatively less salient than the two-referent distractor, simply by virtue of having one rather than two features. Thus, whether our current data signal a true developmental change or simply a limitation of our experimental methods will be a question for future work.

Adult Control Data

Adults performed at ceiling in the online version of our task, choosing the one-feature object 96% of the time in the Label condition (95% CI: 91% - 100%) and 23% of the time in the No Label condition (95% CI: 15% - 30%). This level is substantially higher than that of the older four-year-olds in our study. But we caution against a strong interpretation of this finding. Adults expressed explicit understanding of the pragmatic nature of the task (in the Label condition, 29% of the responses to our debriefing question "What did you think this study was about?" made direct or indirect mention of informativeness or the dichotomy between what is said and what is meant).

In followup work using the same task as we introduced here, we have found greater levels of explicit metacognition when participants perform multiple trials in a row, as in the results reported above. (In contrast, we find no such order effects in data from children). Thus, most of our experiments with adults have relied on asking a single question to each participant (Frank & Goodman, 2012; Vogel, Emilsson, Frank, Jurafsky, & Potts, 2014). In one study that used substantively identical displays, but only asked a single question to each participant, we found levels of implicature that were almost exactly the same as those shown by the 4-year-olds: 75%. Though many factors may affect the strength of adults' implicatures (Goodman & Stuhlmüller, 2013; Degen & Tanenhaus, 2011), we suspect that the level of inference shown by the oldest children in our task here is not anomalous from the perspective of adult judgments.

Discussion

We began by describing preschool children's puzzling difficulties with one type of pragmatic inference: scalar implicature. We then went on to test their ability to make contextually-grounded, ad-hoc implicatures—inferences that do not rely on linguistic scales using quantifiers or modals. Our experiments provide evidence that children by age 3.5, and perhaps even slightly earlier, can make such inferences. These data provide evidence for preschool children's pragmatic capacities and delimit the class of explanations that can account for failure in scalar implicature more specifically.

Together with work by Barner et al. (2011), this finding begins to suggest a possible resolution to the puzzling pattern of failures in scalar implicature experiments: Children are sometimes capable of computing implicatures, but these implicatures are sensitive to the availability of the inferential alternatives. On the classic Gricean account, inferring that "some" means SOME BUT NOT ALL requires considering the counterfactual scenario in which the speaker wanted to talk about ALL and chose the message "all." Barner et al. (2011) argue that it is this computation that proves troublesome for preschool children: they cannot summon ALL (and its matching message "all") to mind as alternatives that are relevant in the pragmatic computation—at least not in time to have this inference inform their judgements.

The prediction of this account is that when the inferential alternatives are more available, the implicature computation should be easier. An item can be more available for a number of reasons. One reason is that the relationship between the target item and its alternative might be well-practiced. For example, numbers seem to elicit strong scalar inferences even for young children (e.g., "two" could mean TWO OR MORE but is effortlessly narrowed to TWO AND NO MORE, presumably because of an inference from the alternative "three"). For children, "three" is highly associated with "two" because of their positions in the highly-practiced count list (cf. Huang, Spelke, & Snedeker, 2013). In contrast, children are not taught to recite the quantifier list "none," "some," "all" (Barner & Bachrach, 2010).[2] Another, perhaps more straightforward case of accessibility is when the alternative interpretations are pictured in the context. In the case of standard scalar implicature tasks, a "some"-consistent display is shown and children are asked to make judgments about it (Papafragou & Musolino, 2003). There is no display showing "all" in this case. In contrast, in referent selection tasks like ours (or like the Miller et al., 2005 study cited above), the alternative interpretations are physically pictured and hence presumably easier to reason about. Thus, across the number and ad-hoc implicature case studies, there is some prima facie support for the "availability of alternatives" hypothesis.

Other aspects of our experimental design likely played a role in younger children's success in our task as well, however. Tasks that ask for truth-value or felicity judgments impose considerable demands on children beyond their comprehension of an utterance. In contrast, referent selection is a task that children are called upon to perform nearly every day of their lives. By its nature it implies that responses are likely to be exclusive, a feature that may also have pushed children to consider the contrasting modes of referring to the different possible targets in our task (e.g. "if he had wanted this one he would have said 'hat,' but instead he said 'glasses.' "). Future work should capitalize on design features of our task to probe further the sources of previous failures in SI (Miller et al., 2005). Nevertheless, the success of four-year-olds and older three-year-olds in our study

[2] We acknowledge that this interpretation of numbers as being pragmatically upper-bounded is controversial, but cite it for the sake of completeness.

suggest that children's difficulty with some scalar implicatures should not be interpreted as a more general difficulty with pragmatic reasoning.

Pragmatic reasoning is a central area in human cognition where language understanding and social cognition come together to enable sophisticated feats of communication. Because of this centrality, the results on children's limited pragmatic abilities present an important experimental puzzle. Our work here places one piece by showing that some pragmatic inferences are within the capabilities of young children.

References

Baldwin, D. A. (1993). Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental Psychology*, *29*, 832.

Barner, D., & Bachrach, A. (2010). Inference and exact numerical representation in early language development. *Cognitive Psychology*, *60*, 40 - 62.

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

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278.

Bloom, P. (2002). Mindreading, communication and the learning of names for things. *Mind & Language*, *17*, 37.

Braine, M., & Rumain, B. (1981). Development of comprehension of "or": Evidence for a sequence of competencies. *Journal of Experimental Child Psychology*, *31*, 46-70.

Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition*, *112*, 337–342.

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 (Eds.), *Bucld 25 proceedings* (p. 157-168). Somerville, MA: Cascadilla Press.

Clark, E. V. (1988). On the logic of contrast. *Journal of Child Language*, *15*, 317–335.

Clark, E. V., & Amaral, P. (2010). Children build on pragmatic information in language acquisition. *Language and Linguistics Compass*, *4*, 445.

Clark, H. (1996). Communities, commonalities, and commication. In J. J. Gumperz & S. C. Levinson (Eds.), *Rethinking Linguistic Relativity.* Cambridge University Press.

Degen, J., & Tanenhaus, M. K. (2011). Making inferences: the case of scalar implicature processing. In *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 3299–3304).

Foppolo, F., Guasti, M. T., & Chierchia, G. (2012). Scalar implicatures in child language: Give children a chance. *Language learning and development*, *8*, 365-394.

Frank, M., & Goodman, N. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998.

Frank, M., Goodman, N., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*, 578.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models* (Vol. 625). Cambridge University Press Cambridge.

Gergely, G., Bekkering, H., & Király, I. (2002). Rational imitation in preverbal infants. *Nature*, *415*, 755.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, *5*, 173–184.

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

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

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

Guasti, M. T., Chierchia, G., Crain, S., Foppolo, F., Gualmini, A., & Meroni, L. (2005). Why children and adults sometimes (but not always) compute implicatures. *Language and Cognitive Processes*, *20*, 667.

Hirschberg, J. L. (1991). *A theory of scalar implicature*. New York: Garland Pub.

Horn, L. R. (1998). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. *Pragmatics*, 383.

Huang, Y. T., & Snedeker, J. (2009). Semantic meaning and pragmatic interpretation in 5-year-olds: Evidence from real-time spoken language comprehension. *Developmental Psychology*, *45*, 1723-1729.

Huang, Y. T., Spelke, E., & Snedeker, J. (2013). What exactly do numbers mean? *Language Learning and Development*, *9*, 105–129.

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

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

Liszkowski, U., Carpenter, M., Striano, T., & Tomasello, M. (2006). Twelve- and 18-month-olds point to provide information for others. *Journal of Cognition and Development*, *7*.

Liszkowski, U., Carpenter, M., & Tomasello, M. (2008). Twelve-month-olds communicate helpfully and appropriately for knowledgeable and ignorant partners. *Cognition*, *108*, 732 - 739.

Matthews, D., Butcher, J., Lieven, E., & Tomasello, M. (2012). Two- and four-year-olds learn to adapt referring expressions to context: Effects of distracters and feedback on referential communication. *Topics in Cognitive Science*, *4*, 184–210.

Matthews, D., Lieven, E., Theakston, A., & Tomasello, M. (2006). The effect of perceptual availability and prior discourse on young children's use of referring expressions. *Applied Psycholinguistics*, *27*, 403-422.

Matthews, D., Lieven, E., & Tomasello, M. (2007). How toddlers and preschoolers learn to uniquely identify referents for others: A training study. *Child Development*, *78*, 1744–1759.

Meltzoff, A. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental psychology*, *31*, 838–850.

Miller, K., Schmitt, C., Chang, H., & Munn, A. (2005). Young children understand some implicatures. In *Proceedings of the 29 th annual boston university conference on language development* (pp. 389–400).

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science*, *13*, 329-336.

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

O'Neill, D. K., & Topolevec, J. C. (2001). Two-year-old children's sensitivity to the referential (in) efficacy of their own pointing gestures. *Journal of Child Language*, *28*, 1.

Onishi, K., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*, 255–258.

Papafragou, A. (2006). From scalar semantics to implicature: children's interpretation of aspectuals. *Journal of Child Language*, *33*, 721.

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

Papafragou, A., & Tantalou, N. (2004). Children's computation of implicatures. *Language*

Acquisition, *12*, pp. 71-82.

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, *18*, 587–592.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition.* Cambridge, Mass.: Harvard University Press.

Stiller, A., Goodman, N. D., & Frank, M. C. (2011). Ad-hoc scalar implicature in adults and children. In *Proceedings of the 33rd annual meeting of the cognitive science society, boston, july.*

Sullivan, J., Davidson, K., & Barner, D. (2011). Children's conversational implicatures. In *Proceedings of the Boston University Conference on Language Development.*

Tomasello, M., & Akthar, N. (1995). Two-year-olds use pragmatic cues to differentiate reference to objects and actions. *Cognitive Development*, *10*, 201.

Vogel, A., Emilsson, A. G., Frank, M. C., Jurafsky, D., & Potts, C. (2014). Learning to reason pragmatically with cognitive limitations. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society.*

Woodward, A. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*, 1–34.