# The trouble with quantifiers: Children's difficulty with "some" and "none"

Alexandra C. Horowitz

Department of Psychology, Stanford University

Rose M. Schneider

Department of Psychology, Stanford University

Michael C. Frank

Department of Psychology, Stanford University

We gratefully acknowledge Tamara Mekler for her assistance in data collection, the staff and families at Bing Nursery School, as well as David Barner for guidance on the Give-Quantifier task.

Address all correspondence to Rose M. Schneider, Stanford University, Department of Psychology, Jordan Hall, 450 Serra Mall (Bldg. 420), Stanford, CA, 94305. Phone: 650-721-9270. E-mail: rschneid@stanford.edu.

## Abstract

Using language requires speakers and listeners to make pragmatic inferences that extend beyond the literal sense of the utterance. For example, adult listeners quickly and easily infer from the utterance "*Some* of the cookies are oatmeal raisin" that the others might not be. On the other hand, children have more difficulty with this statement; children often fail to make scalar implicatures using quantifiers. Children's performance varies tremendously across studies and tasks however, limiting the number of possible direct comparisons between datasets. To address this issue, we designed a novel experimental paradigm, which both minimized task demands for participants, and enabled comparisons across different tasks. In Experiment 1, we used this paradigm to explore children's ability to compute both ad-hoc (contextual) and scalar (quantifier) implicatures, and found that while older 4-year-olds performed at ceiling for ad-hoc descriptions, they still performed poorly with scalar descriptions. Intriguingly, we also found bimodal and correlated performance with the quantifiers "some" and "none". In Experiment 2, we attempted to isolate children's sources of difficulty with these terms by including only scalar trials, and found that while performance increased, it was still low. In Experiment 3, we explored possible sources of developmental difficulty in this task; inhibitory control did not predict the ability to make scalar implicatures, but children who had difficulty with "some" and "none" in an implicature task also had issues with these terms in a quantifier knowledge task. Taken together, our results provide more detailed developmental data on the development of pragmatic implicatures, and suggest that difficulty with scalar implicatures may in fact be rooted in a lack of quantifier knowledge.

Keywords: pragmatics, development

# Introduction

Adult language users effortlessly make inferences that go beyond the literal sense of an utterance. For example, an adult who hears “I ate *some* of the cookies” would expect that I did not eat *all* the cookies. Similarly, an adult listener who hears “I ate the sugar cookies,” would most likely assume that I ate *only* the sugar cookies, and not the other varieties. These two statements use a weaker literal description to imply that a stronger alternative is true. The first statement requires the listener to make what is known as a scalar implicature, which relies on lexical scales such as quantifiers (“some” vs. ”all”) or modals (“possibly” vs. “definitely”) (Horn, 1972). The second statement necessitates an ad-hoc or contextual implicature, in which a stronger description is negated by a contextually weaker description.[1] Previous research on children’s implicature comprehension has suggested that while children are able to make ad-hoc implicatures, they experience difficulty in computing scalar implicatures. Thus, they would have trouble inferring from “I ate *some* of the cookies” that (a few) cookies still remain. Here, we explore the developmental trajectory of children’s ability to make both ad-hoc and scalar implicatures, and investigate factors influencing children’s performance across scalar descriptions.

Children’s processing of scalar implicatures, is an intriguing case study in pragmatic development. In contrast to adults’ spontaneous computation of scalar implicatures along quantifier lexical scales like <SOME, ALL>, children’s performance on these same scales is variable even until fairly late in development (Noveck, 2001). It is possible however that some of the paradigms used to test implicature comprehension in children might obscure their pragmatic abilities, such as ones that require children to make truth judgments for complex propositions. In

[1] While Grice (1975) distinguished “generalized” and “particularized” implicatures, this distinction has been controversial. Here we use “ad-hoc implicature” as a term of convenience to describe contextually-supported inferences while remaining agnostic about the theoretical distinction.

fact, children show a graded pattern of successes and failures across different tasks (Guasti et al., 2005; Papafragou & Musolino, 2003; Papafragou & Tantalou, 2004). For example, five-year-olds asked to rate the felicity of a statement by selecting the magnitude of a reward (rather than making a binary true/false decision) assigned only mid-sized rewards for true but pragmatically odd descriptions (Katsos & Bishop, 2011), suggesting that they do recognize that weak statements are less felicitous than stronger ones. The wide range of methods and measures used to assess children's ability to compute implicatures also does not allow for many direct comparisons in children's performance across different tasks.

In (Table 1), we present a brief review of developmental implicature research. This is a non-exhaustive review, but does capture the wide range of paradigms and dependent variables across studies. In this literature, we found also that most studies targeted relatively narrow age bins (12–18 months, although several are wider, see cf. Papafragou & Tantalou, 2004). In reviewing these differing methods and measures, we observed several variables which may affect children's ability to make implicatures. For example, we found differing syntactic structure in scalar implicature prompts: recent work indicates that the use of the partitive is a particularly strong cue for scalar implicatures (Degen & Tanenhaus, 2014). Further, paradigms also differed in the availability of both lexical and visual alternatives, both of which have been shown to facilitate implicature computation (Stiller, Goodman, & Frank, 2014). In sum, the diversity of measures across these different studies may result in children's inconsistent performance, as many things in context affect the strength of scalar implicatures (Degen, 2015). Across these studies however, children's difficulty with scalar implicatures suggests that their fragile performance might have less to do with general pragmatic knowledge per se, and more to do with their knowledge of particular scales.

The *Alternatives Hypothesis*, proposed by Barner and colleagues (Barner & Bachrach, 2010; Barner, Brooks, & Bale, 2011), posits that children's ability to compute scalar implicatures relies on their recognition of the relevant lexical alternatives (e.g., that use of the weaker term "some" conveys a direct contrast with the stronger alternative "all", thus implying *some but not all*). In other words, children's pragmatic inferences rely on their ability to consider relevant possible alternative word choices that could have been used in place of the ones the speaker chose. So even in supportive paradigms, if children cannot bring to mind "all" when reasoning about "some," they will fail to make an implicature.

Previous research has supported this hypothesis, with children's performance in implicature tasks increasing when they have stronger access to lexical alternatives. Evaluating performance in competitions where alternative outcomes are salient (Papafragou & Musolino, 2003) and using contextually-accessible (ad-hoc) scales (e.g., "the cat and the cow are sleeping" rather than "some animals are sleeping"; Barner et al., 2011) both help preschoolers make implicatures. The Alternatives Hypothesis also predicts interactions between the supportiveness of a task and children's performance. For example, even older three-year-olds show evidence of computing implicatures for ad-hoc contextualized scales when the task is a referential forced choice between possible interpretations (Stiller et al., 2014). Preschoolers also show some preliminary evidence of computing scalar implicatures for quantifiers in a similar forced-choice paradigm, albeit with prosodic support (Miller, Schmitt, Chang, & Munn, 2005). In sum, the Alternatives Hypothesis appears to provide a promising account of the current patterns of preschoolers' successes and failures in pragmatic implicature tasks.

| Study | Scale(s) | Ages | Measure | Sentence Type | Main Finding |
|---|---|---|---|---|---|
| Noveck (2001) | non-necessity–necessity, possibility–impossibility, some–all | 5;1–5;11,<br>7;1–8;0,<br>9;0–9;5,<br>10;0–11;7 | Truth Value Judgment (*Yes, I agree* or *No, I do not agree*) | "Some giraffes have long necks" | Comprehension matters: Children demonstrate pragmatic competence by age 7 in evaluating a variety of plausible and implausible sentences. |
| Papafragou & Mussolini (2003) | some–all,<br>two–three,<br>start–finish | 4;11–5;11 (Study 1)<br>5;1–6;5 (Study 2) | Felicity Judgment (*Did Minnie answer well?*) | "Some of the horses jumped over the fence" (when all of the horses jumped over the fence) | Support matters: Children were more likely to reject infelicitous weak descriptions for numbers, and for all types of weak descriptions in the task with more pragmatic support (informativeness training, context of competition, statements about specific events). |
| Papafragou & Tantalou (2004) | some–all,<br>ad-hoc,<br>encyclopedic | 4;1–6;1 | Felicity Judgment (Decide whether or not to award a speaker a prize) | *Did you color the stars?* "I colored some (when all were colored) | Scales matter: Children mainly withheld prizes for weak descriptions, and at higher rates for ad-hoc trials than other trial types. |
| Guasti, Chierchia, Crain, Foppolo, Gualmini, & Meroni (2005) | some–all | 7;0–7;7 | Truth Value Judgment: *Yes, I agree* or *No, I do not agree* (Studies 1–3), *Did Carolina say the wrong thing?* (Study 4) | "Some giraffes have long necks" (replication of Noveck (2001)), and scene descriptions, e.g. "Some monkeys are eating a biscuit" (when all are) | Context matters: 7-year-olds reliably computer implicatures for "some" after a training that increased their sensitivity to the informativeness of speakers' descriptions (e.g., calling a grape "fruit" instead of "grape") and when contextualized as evaluating a novice speaker novice speaker describing a scene. |
| Miller, Schmitt, Chang, & Munn (2005) | some–all | 4;1–5;5 (Study 1)<br>3;6–5;10 (Study 2) | Direct Instruction Task (Study 1);<br>Picture Matching Task (Study 2) | "Make some faces HAPPY/Make SOME faces happy/Make some HAPPY faces" (Study 1), "Shoe me where Pete made some faces HAPPY/Show me where Pete made SOME faces happy" | Prosody matters: In both tasks (completing the scene or selecting the referent), children reliably identified only a subset of the faces (out of four) when "some" was stressed, but not when it was unstressed. |
| Huang & Snedeker (2009) | some–all,<br>two–three | 5;2–6;1 (Study 1)<br>5;5–6;9 (Studies 2 & 3) | Eye-tracking<br>referent selection | "Point to the girl with some of the socks" (when other girls and boys have shares of socks and soccer balls) | Time scale matters: Across studies, children were delayed in identifying the referent for scalar implicature trials, and accept and overlap between the meaning of "some" and "all." |
| Katsos & Bishop (2011) | some–all,<br>ad-hoc | 5;1–6;3 | Binary Truth Value Judgment (Study 1); Ternary Truth Value Judgment (Study 2); Sentence-to-picture Matching Task (Study 3) | "The mouse picked up some of the carrots" | Measures matter: While children tended to a accept under-informative scalar and ad-hoc descriptions given a binary decision, they showed sensitivity to weaker statements given a ternary choice or picture matching task. |
| Barner, Brooks, & Bale (2011) | some–all,<br>ad-hoc | 4;0–5;0 | Truth Value Judgment | "Are some of the animals sleeping?" (when all are) | Specificity matters: 4-year-olds accept weak ad-hoc and scalar descriptions. When preceded by restrictive "only", they reject ad-hoc descriptions but continue to accept that "only some" can mean *all*. |
| Skordos & Papafragou (2014) | some–all | 4;9–5;8 | Felicity Judgment (*Did the puppet answer well?*) | "Some of the blickets have a crayon" (when all of them do) | Comparisons matter: Children were more likely to reject infelicitous uses of "so" if they first heard "all" falsely refer to quantity (only 3/4 blackouts had crayons), but not is "all" referred falsely to the objects (e.g., "all of the blackest have a scarf"). |

Table 1

*Review of previous literature on children's comprehension of implicatures.*

We designed a simple referent selection task in which children were asked to select which of three book covers they thought the experimenter was describing. Our design allowed us to fully counterbalance the trial types (ad-hoc vs. scalar descriptions crossed with implicature vs. unambiguous control targets) across participants, to examine both within-subject patterns of responses and between-subject developmental patterns, and to reduce the demands of the task by having children select the implied referent among three visual alternatives.

In Experiment 1, we included both ad-hoc and scalar descriptions with implicature and control trials for each. Four–year-olds were strong on ad-hoc trials (similar to previous work, e.g. Stiller et al., 2014), but their performance on scalar implicature trials was very low. In Experiment 2, we ran the same task but omitted ad-hoc trials. We found developmental increases in performance for each trial type, and higher performance on implicature trials for 4-year-olds in this scalar-only version of the task. In both Experiments 1 and 2, we found and unexpected result. Children's pattern of responses on scalar implicature trials was bimodal and strongly correlated with their performance on "none" (scalar control) trials, providing some clues about the factors underlying success in scalar implicatures. In Experiment 3, we explored two alternatives (inhibitory control problems and lack of quantifier knowledge) that could be contributing to children's difficulty with the quantifiers "some" and "none". Overall, our findings suggest that while preschoolers' computation of scalar implicatures can be supported by stronger recognition of the lexical alternatives, that their failure in making these implicatures may be rooted in difficulty using and contrasting quantifiers.

## Experiment 1: Ad-hoc and scalar implicature computation in children

Given the difficulty in equating results on children's computation of implicatures across different methods and paradigms, we created a single task that could be adapted to investigate both ad-hoc and scalar items in one task. This task involved one set of visual stimuli presented in the same order to all participants; however, the particular items (ad-hoc or scalar) queried were counterbalanced across participants. Thus, with one set of visual stimuli we could directly compare children's performance on both ad-hoc and scalar implicatures in a single experimental session. In Experiment 1, we included questions about ad-hoc and scalar implicatures within one session, and found that this paradigm was appropriate for ad-hoc items, but children still experiencesd difficulty computing scalar items.[2] In all scalar items, we used a partitive construction to increase the likelihood of making an implicature (Degen & Tanenhaus, 2014).

### *Methods*

*Participants.* A planned sample of 48 children was recruited from a university preschool. These children were drawn from two age groups: twenty-four 4.0 – 4.5-year-olds (M = 4;2, median = 4.19, SD = 0.14) and twenty-four 4.5 – 5.0-year-olds (M = 4.74, median = 4.73, SD = .16). Two children were excluded from the final sample for not completing the task, and one additional child was excluded due to experimenter error. All children's primary language was English, and no child completed more than one session of the task.

*Stimuli.* Stimuli for all experiments were created to be appropriate for questions about both ad-hoc and scalar implicatures, allowing the experimenter to use one set of stimuli for both kinds of items in one experimental session. The experimental stimuli consisted of a set of printed pictures of three book covers with four familiar items on each cover. In each trial, one book cover contained four items of the same kind (e.g., four cats), another book cover contained four items of another kind (e.g., dogs), and the final cover contained two items of a new set and two items

[2] All stimuli, data, and analyses are available in a public repository at `http://github.com/rosemschneider/si_paper`.

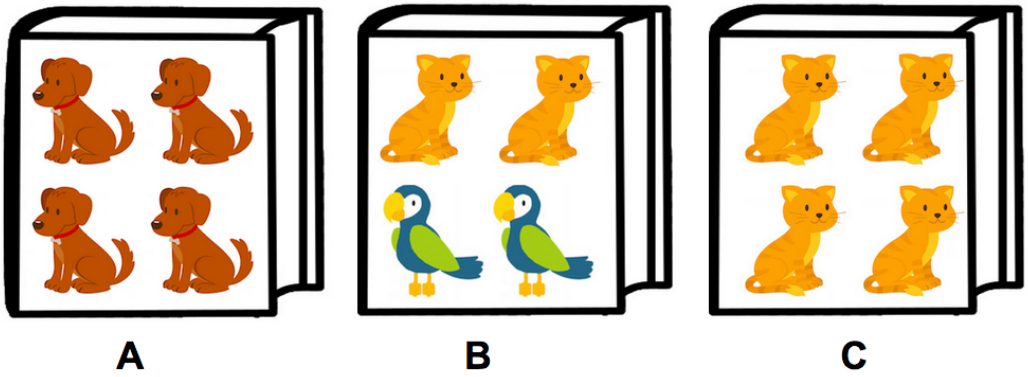

*Figure 1.* Example trial stimuli used in all experiments. Children received a clue from the experimenter about which book she had in mind and responded based solely on the clue; this was either an ad-hoc or a scalar description of a book with either an unambiguous or implicature target.

repeated from one of the other book covers (e.g., two birds and two cats). An example of the stimuli can be seen in Figure 1. All items on the book covers were familiar to children, and were able to be identified. All participants saw the same book covers in the same order.

*Procedure*. Participants were tested in individual sessions in a quiet room at their nursery school. The experimenter introduced the study as a guessing game, and explained that the child would receive a hint about which book cover the experimenter had in mind. In the instructions for the task, the experimenter emphasized that the child would only receive one clue about what book the experimenter was describing, and they had to use that clue to make their decision. All participants saw the book covers in the same order; however, three scripts with both ad-hoc and scalar trials were counterbalanced across participants. A breakdown of trial types and sample scripts can be seen in Table 2.

| Condition | Trial type | # trials, Expt. 1 | # trials, Expts. 2 & 3 | Statement: "On the cover of my book, ..." | Target |
|---|---|---|---|---|---|
| Scalar | implicature | 4 | 6 | "...some of the pictures are cats" | B |
| | all | 2 | 6 | "...all of the pictures are cats" | C |
| | none | 2 | 6 | "...none of the pictures are cats" | A |
| | unambiguous 'some' | 2 | | "...some of the pictures are birds" | B |
| Adhoc | implicature | 4 | | "...there are cats" | C |
| | distractor | 2 | | "...there are dogs" | A |
| | comparison | 2 | | "...there are birds" | B |

Table 2
*Study designs for Experiments 1, 2, and 3, using script examples for the trial set pictured in Figure 1.*

Prior to the test trials, children were familiarized to the task with a practice trial with three book covers, each displaying a single unique and familiar item. During the practice trial, the experimenter told the child "On the cover of my book, there's a TV," which corresponded to the middle book cover. After children had successfully completed the practice, they saw 18 test trials with stimuli of books containing sets of familiar items. At the start of every trial, the experiment would provide the child with either an ad-hoc or scalar description of one of the books and instruct the child to point to the book she was describing. If the child pointed to more than one book, or the response was otherwise ambiguous, the experimenter emphasized again that she was talking about just one book, and that they should choose the single book she was describing.

In ad-hoc trials (eight total), the experimenter's descriptions of the target book used names of the pictured objects, providing contextual support for the target. Ad-hoc control trials referred to an unambiguous target (e.g., "On the cover of my book, there are dogs" in Figure 1), while implicature trials required the child to reason about the speaker's meaning given the ambiguous utterance (e.g., “On the cover of my book, there are cats”, which could refer to either the book containing only cats or the book containing cats and birds). In these critical trials, children had to understand that the speaker could potentially be talking about either the book with four or two of the named object, but that by opting to describe only one kind of object she was referring to the cover with four of the same object; otherwise, she would have mentioned both kinds of objects, or the ones unique to that cover (i.e., birds).

In scalar trials (ten total), the experimenter described the target book with quantifiers. For scalar items, control trials referred to unambiguous targets with the quantifiers “all” and “none” (e.g., “On the cover of my book, *all/none* of the pictures are cats”) or an unambiguous referent of “some” (e.g., “On the cover of my book, *some* of the pictures are birds”). On critical scalar implicature trials, the experimenter used the weak quantifier “some” to reference the item pictured across two book covers (e.g., “On the cover of my book, *some* of the pictures are cats”). These trials required the child to reason that because the speaker used the weak quantifier “some”, she must be referring to the book picturing only two of the named target, or else she would have used the stronger quantifier “all”.

All participants saw image sets in the same order; however, these image sets were counterbalanced for target location across the three scripts. Description condition and trial-type were further randomized across participants, and were spaced to avoid immediate repeat trial types. Children did not receive feedback after the test trial.

### *Results*

Children’s accuracy on all trial types is plotted in Figure 2. In implicature trials, children’s performance was coded as correct if they selected the image consistent with either the ad-hoc or scalar inference. Children were at ceiling making ad-hoc implicatures, which is consistent with

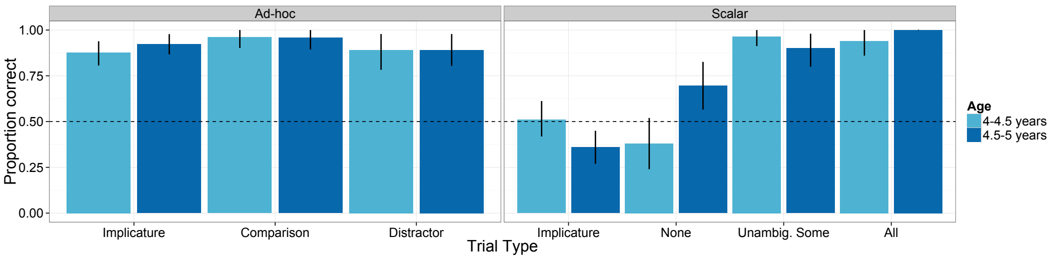

*Figure 2*. Proportion of correct responses by each age group across all trial types and split by implicature type. Error bars show 95% confidence intervals computed by non-parametric bootstrap.

previous research suggesting that children are able to succeed in making such implicatures when they have access to the relevant lexical alternatives (Stiller et al., 2014). Children's performance across ad-hoc trials provides strong evidence that our novel paradigm is an appropriate measure for such items.

In contrast to their success in making ad-hoc implicatures, children struggled in scalar trials. Although they succeeded in "all" and "unambiguous some" trials, children performed at chance on "some" trials (4–4.5-year-olds: $t(22) = .4$, $p = .72$; 4.5–5-year-olds: $t(24) = -2$, $p = .06$) and at chance on "none" trials (4–4.5-year-olds: $t(24) = -.9$, $p = .4$; 4.5–5-year-olds: $t(24) = 1.6$, $p = .13$).

In a planned analysis, we ran a logistic mixed effects model, predicting a correct response as an interaction of age, condition (ad-hoc or scalar) and trial type (implicature or control), with random effects of participant and trial type. We found that performance was slightly lower for scalar trials than ad-hoc trials ($\beta = -8.02$, $p = .09$), and that there was a significant interaction between condition and trial type, such that performance was significantly worse on scalar implicature trials ($\beta$= 16.45, $p$ <.02). We also found a significant 3-way interaction between condition, trial type, and age, such that performance on scalar implicature trials decreased with age ($\beta$ = –4.16, $p$ <.01). There were no significant effects of adding trial order (trials in the first half vs. second half of the experiment), indicating that performance did not change throughout the course of the experiment.

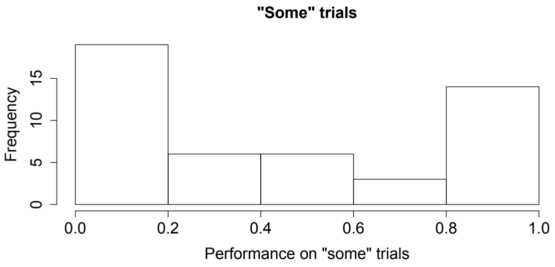

*Figure 3*. Histogram of performance on "some" trials

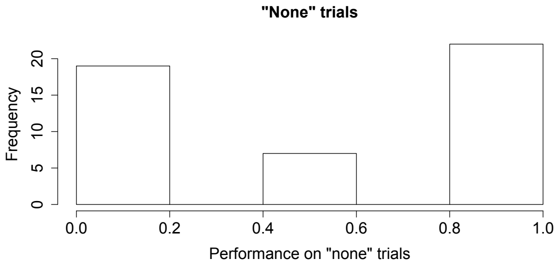

*Figure 4*. Histogram of performance on "none" trials

In post-hoc analysis of the data, we found an unpredicted consistency in performance on "some" and "none" (Figures 3 and 4).[3] To examine their patterns of responses more closely, we ran Hartigan's dip test and found significant bimodal distributions for both *some* ($D$ = .15, $p$ <.0001) and *none* ($D$ = .20, $p$ <.0001). This suggests children did not respond at chance in scalar trials, but consistently either correctly or incorrectly. Additionally, children's success on *some* and *none* trials was highly correlated ($r$ = .47, $p$ <.001), such that children who performed better on some trials also tended to perform better on none trials (Figure 5). Performance on *none* and *all* trials ($r$ = .11, $p$ = .45) and *some* and *all* trials ($r$ = .01, $p$ = .95) was not correlated.

[3] We also observed this bimodal performance in a pilot sample (N = 23), and found that it replicated in Experiment 1.

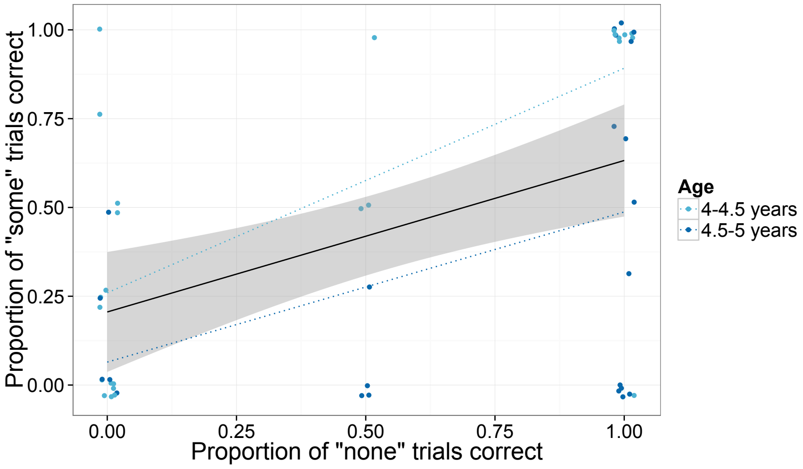

*Figure 5*. Scatterplot relating individuals' performance on "some" and "none" trials per age group in Experiment 1. The aggregate trend is plotted in black along with its 95% confidence interval, and trends for individual age groups are shown by dotted lines. Points are jittered slightly to avoid over plotting

### *Discussion*

The results of Experiment 1 indicated that while children were easily able to make ad-hoc implicatures in our task, they had difficulty making scalar implicatures. This pattern of performance was puzzling, given both our efforts to reduce task demands and children's striking success in ad-hoc trials. Despite having access to both visual alternatives (the three selection choices) within each trial and lexical alternatives across all trials, children were still at chance in making scalar implicatures in our task.

Even more intriguing was the unexpected developmental change we observed on *none* trials. We included "none" as an unambiguous control quantifier, but found that children performed at chance for this scalar term as well. These results are supported by previous work suggesting that even older preschoolers struggle with negation occurring in contexts without pragmatic support (Nordmeyer & Frank, 2014). Given children's difficulty with "none" and the strong positive correlation between *some* and *none* trials, it is possible that making implicatures necessitates some familiarity with both ends of the quantifier scale (*none – some – all*). An inability to use and contrast relevant alternatives along the lexical scale <SOME, ALL> might lead to failure in computing scalar implicatures. In addition to understanding the extremes of the quantifier scale, two other possibilities – namely, lack of quantifier knowledge, and inhibitory control – may also account for children's decreased performance on scalar items.

However, we wondered if these results were the effect of including both ad-hoc and scalar quantifier descriptions within one experimental session. It is possible that children's success on ad-hoc trials may have to a misinterpretation of a scalar description (e.g., "On the cover of my book, some of the pictures are cats") to an ad-hoc one ("On the cover of my book, there are cats"). Presenting scalar descriptors in the same experimental session as other relevant and felicitous descriptions inhibits scalar implicature comprehension, even in adults, cf.(?, ?).

To further explore the possibility that children's performance on scalar trials might have been influenced by ad-hoc trials, we removed all ad-hoc trials from our task, and ran a scalar-only version of the the study.

## Experiment 2: Isolating scalar implicatures

In Experiment 2, we pursued the possibility that children might be failing to make scalar implicatures as a result of competing ad-hoc descriptors in the same experimental session. Additionally, we expanded our target age-range to 3–5 years in order to more fully explore the developmental trajectory across implicature and control trials.

*Methods*

*Participants*. We recruited a new sample of 50 participants from a university preschool: twelve 3.0–3.5-year-olds (M=3;4, median = 3.35, SD = 0.1), twelve 3.5–4.0-year-olds (M=3;8, median = 3.67, SD = .12), fourteen 4.0–4.5-year-olds (M=4;3, median = 4.24, SD = .1), and twelve 4.5–5.0-year-olds (M=4;8, median = 4.63, SD = .15). One additional child was excluded for stopping the task early.

*Stimuli*. Stimuli were identical to Experiment 1. The only changes made in experimental protocol were to the scripts; all 18 test trials were converted to quantifier descriptions (Table 2). In Experiment 2, the 18 test trials consisted of six control *all* trials (e.g., "On the cover of my book, *all* of the pictures are cats"), six *none* trials (e.g., "...*none* of the pictures are cats"), and six scalar implicature trials ("...*some* of the pictures are cats"). We removed the unambiguous *some* trials to more effectively counterbalance; in "some" trials, the quantifier always referenced the item pictured across two book covers (e.g., in Figure 1, children heard references to *none, some,* or *all* cats). As in Experiment 1, image sets were presented in a fixed order, counterbalanced for both target location and triad order. Participants were randomly assigned to one of three scripts, with a pseudo-randomized trial order such that every book set was referred to by each quantifier type, and the same trial type never immediately repeated.

*Procedure*. The procedure was identical to Experiment 1.

*Results*

In Experiment 2, children's performance over all trial types increased with age (Figure 6). Performance was highest in *all* trials, with all age groups significantly above chance ($p < .05$ for all tests). However, performance was still low in both *none* and *some* trials, with only the

4.5–5-year-olds performing above chance for *none* trials ($t(11) = 3.09$, $p < .01$) and only marginally above chance in *some* trials ($t(11) = 1.85$, $p < .09$). Children's performance in Experiment 2 was not significantly different than in Experiment 1, in independent sample t-tests by trial type between age groups in both experiments ($p > .09$ for all tests). All of these tests were relatively low in power due to the small number of individuals in each bin, however.

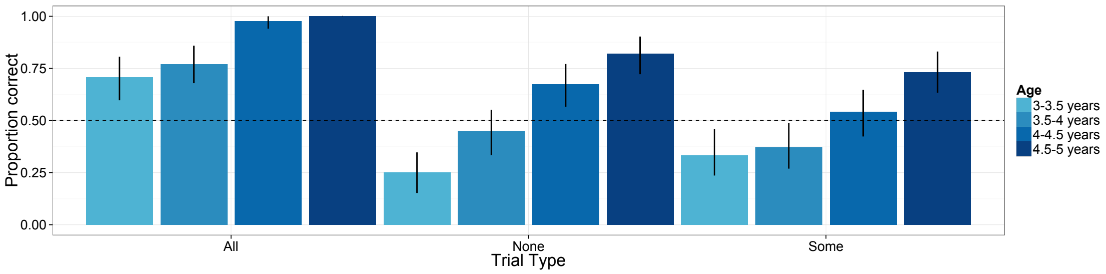

*Figure 6*. Proportion of correct responses by each age group across all scalar trial types. Error bars show 95% confidence intervals computed by non-parametric bootstrap.

To aggregate across groups, we ran a planned logistic mixed effects model, predicting correct responses as an interaction of age and trial type (*all, some*, or *none*), with random effects of trial type by participant. The only significant effect that emerged was age, such that performance increased across trials as children got older ($\beta = 20$, $p < .001$). Adding trial order (first or second half of the experiment) to the model did not interact with any of the variables, indicating the performance did not change over the course of the experiment. We suspected that this lack of a conviction effect was due to individual variability, such as in Experiment 1.

Consistent with the findings from Experiment 1 model, we ran Hartigan's dip test and again found significant bimodal patterns of responses for both *some* ($D = .12$, $p < .0001$) and *none* ($D = .15$, $p < .0001$) trials. Once again, these trial types were highly correlated with one another ($r =$

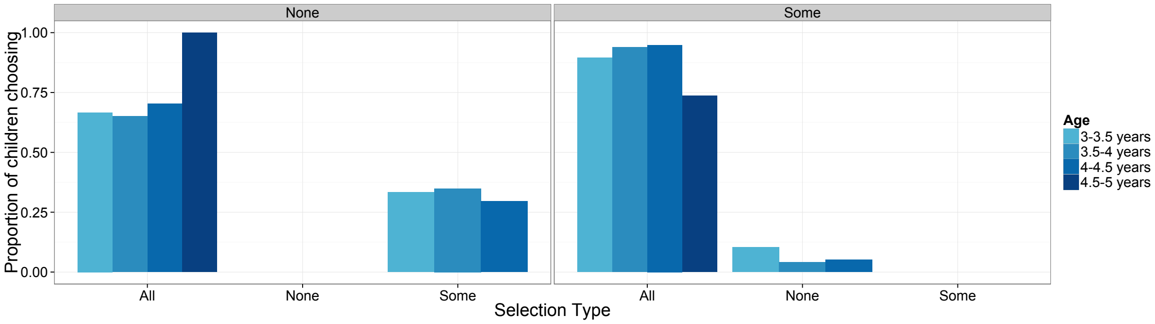

*Figure 7*. Children's selections on incorrect trials, faceted by trial type ("some" and "none"). Graph shows the proportion of children choosing an incorrect alternative.

.52, $p$ <.001). Thus, as an exploratory analysis, we ran another version of the mixed effects model removing the random effect of trial type, as this model does not find trial type effects due to correlation observed between trial types.[4] In addition to a main effect of age ($t$ = 1.88, $p$ <.01), this model a conditional effect: *some* trials were lower than *all* trials ($t$ = –7.69, $p$ <.01), and marginally reduced from *none* trials ($t$ = 3.03, $p$ = .09). It also showed interactions between trial type and age, such that there was a greater difference between younger children's performance on *some* and *all* trials ($t$ = 2.84, $p$ <.001), and *some* and *none* trials ($t$ = 0.90, $p$ = .05). Overall, we observed large individual variability in children's performance, with mean trends of children struggling with the quantifiers "some" and "none" in relation to "all."

In a further exploratory analysis, we were interested in the particular kinds of errors that children made in "some" and "none" trials (Figure 7). We found that on these trials children matched the noun on both kinds of trials, and chose the "all" option more frequently, indicated

[4] Model formula: `(correct ∼ trial type * age + (1 | subject))`

some evidence for either listening through the quantifier (i.e., an issue of inhibitory control), or misunderstanding what the quantifier meant (a lack of quantifier knowledge).

### *Discussion*

In Experiment 1, we observed success in children's computation of ad-hoc, but not scalar implicatures. To explore whether children's performance in making scalar implicatures was hindered by the presence of both ad-hoc and scalar items in the same session, we excluded ad-hoc items, and replaced them with scalar descriptions. When faced with only scalar implicatures, children's performance was numerically (although not significantly) better than in Experiment 1, with success in making a scalar implicature positively correlated with age. We still observed low performance in *some* and unambiguous *none* trials. Additionally, we again found bimodal and correlated patterns of responses in these two trials, with children making errors in i"some" trials also failing on "none" trials.

The results of Experiment 2 indicated that children have difficulty with making scalar implicatures beyond dealing with competing contextual descriptors, but the cause of this developmental change is not clear and individuals differ substantially. One possible explanation for this effect is that children's knowledge of the full quantifier scale is not yet adult-like in their preschool years, and the meanings of these scalar terms may yet be established. Another possibility is that children's performance might suffer on these task from developing inhibitory control, with children not being able to overcome the impulse to select the target noun, regardless of the quantifier used. We explore these two alternatives in Experiment 3, and include measures of both inhibitory control and quantifier knowledge in the same session.

## Experiment 3: Inhibitory control and quantifier knowledge measures

In an effort to explore the particular sources of difficulty for children in making scalar implicatures, in Experiment 3 we designed an individual differences paradigm. We supplemented our implicature task with an inhibitory control task, the Dimensional Change Card Sort (DCCS) (Zelazo, 2006), and a quantifier-knowledge task, Give-Quantifier (Barner, Chow, & Yang, 2009) in a within-subjects design. The DCCC is a standard executive function measure that requires children to shift tasks midway through the task (e.g., sorting cards based on shape rather than color). Children's performance in the DCCS (i.e., their ability to task switch) is a reliable measure of their inhibitory control (Zelazo, 2006). The Give-Quantifier task is a productive measure of children's quantifier knowledge, asking them to give a some number of items in response to a quantifier prompt.

### *Methods*

*Participants.* We recruited a new planned sample of 72 children from a university preschool; this sample was selected to have 80% power to detect correlations of *r* ¿.3. Once again, we included children from 3–5 years: twenty-one 3–3.5-year-olds (M = 3.27, median = 3.28, SD = 0.15), fifteen 3.5–4-year-olds (M = 3.79, median = 3.82, SD = 0.15), nineteen 4–4.5-year-olds (M= 4.17, median = 4.11, SD = 0.13), and seventeen 4.5–5-year-olds (M = 4.71, median = 4.68, SD = 0.14). Twelve children additional were recruited but excluded from the final sample for having participated in either Experiments 1 or 2. Nine children asked for a break, and completed one of the three tasks in a subsequent testing session.

*Stimuli.* The stimuli for the implicature task were identical to Experiment 2. Our measure of inhibitory control was the Dimensional Change Card Sort (DCCS, (Zelazo, 2006)), with materials drawn from that study. The 14 laminated sorting cards used in the study consisted of 7 red rabbits and 7 blue boats. These cards were put into plastic sorting trays, one with a red boat, and the other with a blue rabbit. To assess quantifier knowledge, we used the Give-Quantifier task (Barner et al., 2009). Stimuli for this task consisted of three different sets of plastic fruits (8 oranges, 8 bananas, and 8 strawberries) and a red plastic plate. These fruits were grouped together by kind at the start of each trial.

*Procedure*. The procedure for our implicature task was identical to Experiment 2. Task order was counterbalanced across participants, and individual scripts for each task were also counterbalanced to avoid any possibility of order effects. The tasks were done in a small room apart from the main classroom in individual sessions. The experimenter asked the child before the start of every task whether she would like to play the game or return to the classroom.

We drew our protocol for DCCS directly from the original methods paper (Zelazo, 2006). In running the Give-Quantifier task, we followed the methods of the original paper (Barner et al., 2009), with the exception of the quantifiers used in the task (*some, all, none* and *most*). The experimenter used the partitive construction and prosodically emphasized the quantifier across all trials (e.g., “Can you put *all* of the bananas into the circle?”). Quantifiers were presented in two different orders between participants, and fruit-quantifier pairings were quasi-randomized such that the same pairing was not repeated within a session. If the child requested clarification, the experimenter repeated the prompt, and added that the child should put however many pieces of fruit she felt should go on the plate.

In coding the results of the Give-Quantifier task, we relied on the original coding scheme (Barner et al., 2009); *all* and *none* trials were coded as correct for 8 and 0 pieces of fruit given respectively, whereas *some* trials were coded as correct if the child gave between 2 and 7 pieces,

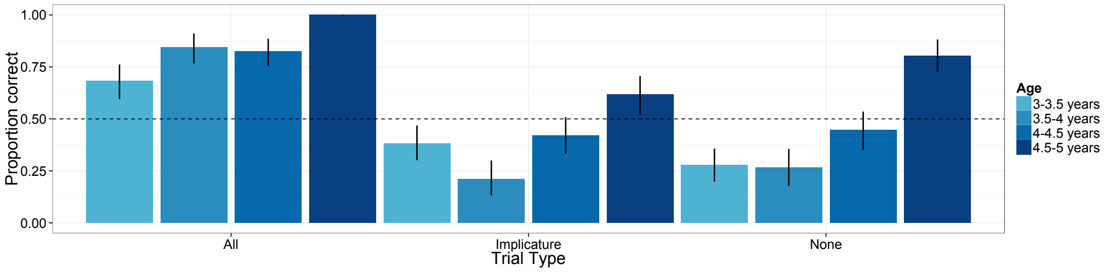

*Figure 8*. Proportion of correct responses by each age group across all scalar trial types. Error bars show 95% confidence intervals computed by non-parametric bootstrap.

and *most* trials were correct if the child gave between 5 and 7 pieces.

### *Results*

We again replicated children's performance on our implicature task (Figure 8); independent sample t-tests between performance in Experiments 2 and 3 within age-groups did not yield any significant differences ($p > .1$ for all tests) except for 4–4.5-year-olds' performance on "all" trials, which was significantly lower in Experiment 3 ($t(31)$, $p < .05$). Once again, we found that performance increased with age, with the 4.5–5-year-olds the only group performing significantly above chance on *none* trials, ($t(16) = 3.75$, $p < .002$), but still performing near chance on *some* trials ($t(16) = 1.18$ $p = .26$). As in Experiment 2, we ran Hartigan's dip test in a post-hoc analysis and again found a significant bimodal distribution of performance in both *some* ($D = .07$, $p = .002$) and *none* trials ($D = .15$, $p > .0001$). Performance with these two quantifiers was also significantly positively correlated ($r = .4$, $p < .001$). Because children's performance was so highly correlated with age however, we ran a partial correlation controlling for age, and found

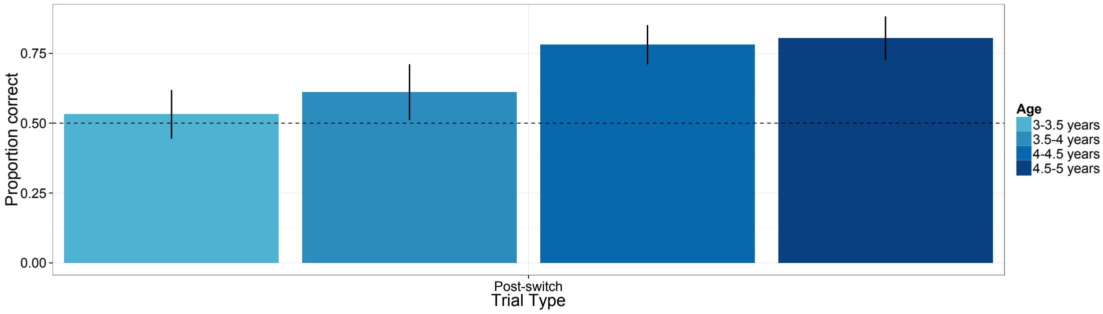

*Figure 9.* Proportion of correct responses by each age group in post-switch trials of DCCS. Error bars show 95% confidence intervals computed by non-parametric bootstrap.

these trial types were still significantly correlated ($r$ = .3, $p$ < .01).

We next turned to whether children's lower and correlated performance on *some* and *none* trials was a result of difficulty with inhibitory control. Accuracy on the DCCS is plotted in Figure 9. This task was significantly correlated with age ($r$ = .28, $p$ = .018), with only 4–5-year-olds performing significantly better than chance (4–4.5-year-olds: $t$(18) = 3.05, $p$ < .01; 4.5–5-year-olds: $t$(16) = 3.31, $p$ <.01). After controlling for age, we did not find any significant correlation between inhibitory control and performance on either *some* trials ($r$ = .13, $p$ = .26) or *none* trials ($r$ = - .01, $p$ = .93) in our implicature task.

Next, we turned our attention to the Give-Quantifier task. Children's performance on this task was very similar to performance on our implicature task, with all age groups performing at ceiling for the prompt *all*, and only older children succeeding on with the prompts *most, none* and *some*. Figure 11 shows the breakdown of how children responded to these scalar terms. We collapsed across all age groups and found a significant bimodal distribution of responses with

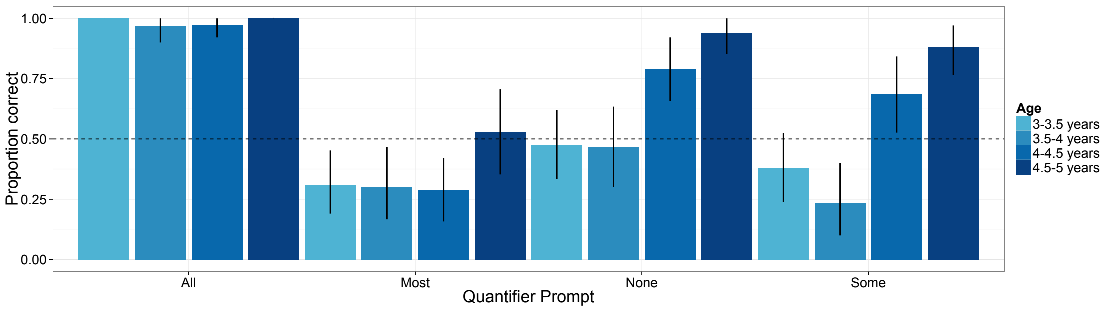

*Figure 10.* Proportion of correct responses by each age group for quantifier prompts *all, most, none* and *some*. Error bars show 95% confidence intervals computed by non-parametric bootstrap.

Hartigan's dip test in both "some" ($D$ = .19, $p$ < .0001) and "none" trials ($D$ = .15, $p$ < .0001). In an exploratory analysis, we ran dip tests by age group, and found that this effect was primarily driven in "some" trials by 3–3.5-year-olds ($D$ = .14, $p$ < .0004) and 4–4.5-year-olds ($D$ = .13, $p$ < .004), and in "none" trials by 3–3.5-year-olds ($D$ = .21, $p$ < .0001) and 3.5–4-year-olds ($D$ = .2, $p$ < .0001). Overall, we found that younger children showed a bimodal pattern of response to the prompt *none*, with the majority of children giving 0 or 8 items in these trials, gradually shifting to a correct response by 4.5–5 years. Similar to performance with *some* trials in our implicature task, we found a large proportion of younger children giving all 8 objects in response to the prompt *some*, and the oldest age groups giving a more adult-like response. In a partial correlation controlling for age, we found that performance in *none* and *some* trials was significantly correlated ($r$ = .61, $p$ <.0001).

In a further exploration of the relationship between quantifier knowledge and scalar implicature performance, we examined both the particular kinds of errors that children made

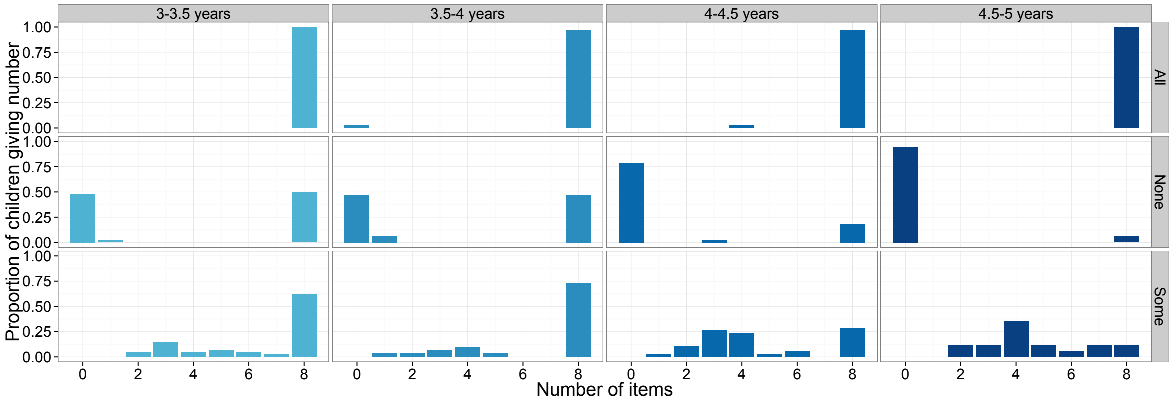

*Figure 11*. Proportion of children giving numbers of items faceted by age group and quantifier prompts *all, most, none* and *some*.

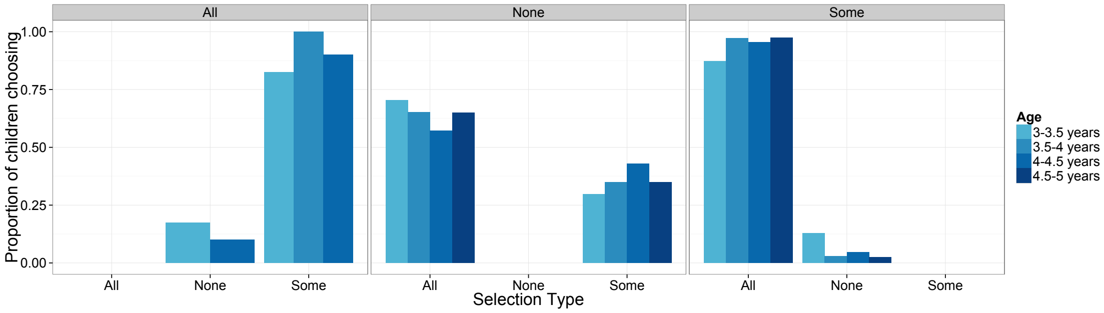

*Figure 12*. Proportion of children choosing an alternative on incorrect trials, faceted by trial type (*some, none*) and split by age group.

across both tasks, and how they were related. Figure 12 shows the breakdown of children's performance in our Scalar Implicature task on incorrect trials. As in Experiment 2, children seemed to make their selections based solely on the target noun, regardless of the quantifier used. This result closely mirrors children's performance in our Give-Quantifier task (Figure 11), in which younger children responded in a bimodal fashion for the items *some* and *none*.

In partial-correlations controlling for age, we found that performance on our Give-Quantifier and implicature tasks were correlated: children who tended to struggle with scalar terms in the context of implicatures also had lower performance when asked to produce a number of items in response to a quantifier prompt, specifically on *some* ($r$ = .27, $p < .02$) and *none* trials in both tasks $r$ = .52, $p < .0001$). We did not find a significant correlation with performance on *some* implicature and *none* quantifier trials ($r$ = .18, $p$ = .14), although we did find a relation between performance on *none* implicature and *some* quantifier trials ($r$ = .35, $p$ = .002).

*Discussion*

In Experiment 3, we turned our attention to investigating potential factors behind children's difficulty with making scalar implicatures. We combined two tasks targeting specific hypothesized areas of difficulty (inhibitory control and lack of quantifier knowledge) with our implicature paradigm in a within-subjects design to explore the relationship amongst these abilities. While we found that younger children did have difficulties in our inhibitory control task, we did not find a significant relation between performance in this task and our other tasks on the basis of inhibitory control, at least at the .3 level that we had power to detect.

We did find that children's performance on the Give-Quantifier and Implicature tasks were related however, and children who failed with scalar items tended to do so across both tasks. This result indicates that processing quantifiers is difficult for children even removed from making implicatures. While it is clear that children struggle with quantifier comprehension, the source of this developmental difficulty is less clear. It is still possible that children's definitions of these scalar terms are not solidified in the pre-school years, and that successfully understanding *some* necessitates understanding both *none* and *all*. It is also possible that the pragmatic contexts in which these scalar items occur are particularly difficult to process.

## General Discussion

We designed a simple task to investigate patterns of pragmatic development both within- and between-subjects. We minimized task demands by asking participants to select the speaker's intended referent from among three visual alternatives. In Experiment 1, we replicated the finding that preschoolers can compute ad-hoc implicatures, though we found poor performance on scalar implicatures. In Experiment 2, preschoolers' comprehension of all scalar quantifier terms in the task increased with age, and removing the ad-hoc trials increased older children's performance on scalar implicature. In Experiment 3, we explored two possible sources of difficulty in implicature computation, inhibitory control and quantifier knowledge, and found that even when removed from an implicature context, quantifier comprehension is difficult for children. Our findings suggest that 4-year-olds are able to compute scalar implicatures with support from contextual cues, but their performance is fragile and may be inhibited by an unestablished quantifier scale.

Our work contributes to the existing literature in a number of ways. First, it offers a novel paradigm that is less complicated than many other implicature tasks, leading us to feel more confident that our results reflect children's true sensitivity rather than inadvertent task demands. Each test set remained visible to children, and they were merely asked to select which picture they thought was the referent corresponding to the speaker's description. Second, the relatively high number of trials for each participant both helped strengthen our analytical power and also offered the possibility for children to identify lexical alternatives as the study progressed. Third, we were able not only to compare performance across age groups, but also to examine individual patterns of responses across the different trial types. This design helped us determine that preschoolers' performance on scalar implicature trials was bimodal and highly related to their performance on *none* trials, which we would have been unlikely to uncover in a purely between-subjects implicature design without controls. Finally, we were able to test two hypotheses about the sources of children's difficult with scalar implicatures, and rule out inhibitory control as a reason for failure.

Our findings also provide support for the Alternatives Hypothesis (Barner & Bachrach, 2010; Barner et al., 2011). First, our ad-hoc trials in Experiment 1 show that preschoolers had no difficulty generally making inferences about contextual descriptions when alternative nominal descriptions were obvious from the context. Performance on scalar trials also appeared to be related to the recognition of a broad set of lexical alternatives, due to both preschoolers' increasing ability to compute scalar implicatures with age (presumably a proxy for familiarity with scalemates) and due to the difference in performance across Experiments 1 and 2. These results support the idea that children's ability to compute implicatures relates to their ability to reason about what other possible utterances a speaker could have used instead. On the other hand, one pattern in our data was more difficult to reconcile with the Alternatives Hypothesis: Children's performance did not change over the course of either experiment. We had expected that, if children's difficulties with scalar implicature were due to a lack of recognition of the contrastive relationship between "some" and "all," that this relationship would be revealed by the two words' consistent use in contrasting references over the course of the many trials that each child completed (Skordos & Papafragou, 2014). The lack of trial order effects we observed could indicate that children in our task did not yet have strong enough comprehension of these terms for contrastive use to matter, or alternatively that our referent-selection task eliminated the problem of summoning the contrasting term to mind and instead foregrounded some other inferential challenge (perhaps that of inhibitory control).

The correlated responses for“some” and “none” trials in both Experiments 1 and 2 presented a particularly interesting puzzle, and suggested that children’s difficulty with these quantifiers could have a common root, either in quantifier knowledge or inhibitory control. *None* is not typically considered part of the same Horn scale as *some* and *all* (because “all” entails *some*, but “some” does not entail *none*—and in fact entails the opposite), but it is nonetheless a lexical contrast along the same quantifier scale. One possibility is that children’s knowledge of the whole quantifier scale plays a role in scalar implicature, though knowledge of logically-false alternatives is not involved in the computations outlined by most theoretical accounts (e.g. Barner et al., 2011). Another is that performance on *none* and *some* trials may be correlated because both scalar implicature and negation comprehension might require inhibiting another response—the positive alternative in the negative case, and the stronger alternative in the implicature case. We explored these two alternatives in Experiment 3, and found support for the possibility that quantifier knowledge, rather than inhibitory control, drives this correlation between children’s responses with the quantifiers “some” and “none.”

Several limitations in the current study provide direction for further work. First ,we did not record children’s response time on the Scalar Implicature trials. It is possible that a measure of inhibitory control on this task (response latencies) might yield more information about the particular processing and inhibitory demands that these implicatures entail. Additionally, our quantifier knowledge measure (Barner et al., 2009) may present many of the same pragmatic hurdles as our implicature task. For example, the syntactic construction of prompts in Give-Quantifier are very similar (e.g., "Can you put *some* of the oranges on the plate?" vs. "On the cover of my book, *some* of the pictures are oranges.") Future work will pursue response latencies as an alternative measure of inhibitory control in addition to exploring to what extent children's difficulty with *some* and *none* are the result of challenging pragmatic contexts. While our results indicate that children's failure in scalar implicature tasks may have a root in absent quantifier knowledge, it is still very likely that the sources of developmental difficulty with implicatures come from several areas.

In sum, our work suggests that children can draw implicatures based on some lexical choices—such as in the case of ad-hoc implicatures—but they still struggle with quantifier-based scalar implicatures until relatively late. This difficulty is not confined to making scalar implicatures, but extends to children's knowledge of the quantifiers *some* and *none*. The particular trouble with these two quantifiers seems to be rooted in incomplete quantifier knowledge rather than inhibitory control. Once children understand the full quantifier scale, they can recognize the relevant category contrasts intended and compute implicatures from these quantifier word choices.

## References

Barner, D., & Bachrach, A. (2010). Inference and exact numerical representation in early language development. *Cognitive Psychology*, *60*(1), 40-62.

Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in childrens pragmatic inference. *Cognition*, *118*(1), 84–93.

Barner, D., Chow, K., & Yang, S. (2009). Finding one's meaning: A test of the relation between quantifiers and integers in language development. *Cognitive Psychology*, *58*(2), 195-219.

Degen, J. (2015). Investigating the distribution of some (but not all) implicatures using corpora and web-based methods. *Semantics and Pragmatics*, *8*(11), 1-55.

Degen, J., & Tanenhaus, M. K. (2014). Processing scalar implicature: A constraint-based approach. *Cognitive Science*, *39*(4), 667-710.

Grice, H. (1975). *Logic and conversation*.

Guasti, M., 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*(5), 667-696.

Horn, L. (1972). The semantics of logical operators in english. *Unpublished doctoral dissertation, University of California Los Angeles*.

Katsos, N., & Bishop, D. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition*, *120*(1), 67-81.

Miller, K., Schmitt, C., Chang, H., & Munn, A. (2005). Young children understand some implicatures..

Nordmeyer, A., & Frank, M. (2014). The role of context in young children's comprehension of negation. *Journal of Memory and Language*, *77*, 25-39.

Noveck, I. (2001). When children are more logical than adults: experimental investigations of scalar implicature. *Cognition*, *78*(2), 165-188.

Papafragou, A., & Musolino, J. (2003). Scalar implicatures: experiments at the semantics-pragmatics interface. *Cognition*, *86*(3), 253-282.

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

Skordos, D., & Papafragou, A. (2014). *Scalar inferences in 5-year-olds: The role of alternatives.* 38th Annual Boston University Conference on Language Development: Cascadilla Press.

Stiller, A., Goodman, N., & Frank, M. (2014). Ad-hoc implicature in preschool children. *Language Learning and Development*, *11*(2), 176-190.

Zelazo, P. D. (2006). The dimensional change card sort (dccs): a method of assessing executive function in children. *Nature Protocols*, *1*, 297-301.