The trouble with quantifiers: ~~Explaining~~ Exploring children's deficits in scalar implicature

Abstract

~~Scalar implicatures (SI) require a listener to make pragmatic inferences that extend~~ We routinely use the context of utterances to infer a meaning beyond the literal ~~meaning of an utterance~~ semantics of their words (e.g., inferring from "I ate some of the ~~cookies" that~~ cookie" I ate some, but not all). ~~Adults reliably make this pragmatic inference but childrengenerally have more difficulty. In a supportive paradigm, three– to five–year–olds~~ We contrasted children's (N=~~171) can compute~~ 209) comprehension of scalar implicatures with quantifiers and ad-hoc implicatures that relied only on contextual details. Four– to five–year-olds computed ad-hoc, but not scalar, implicatures (Experiment 1)~~; implicature computation and successful comprehension of~~ . Unexpectedly, performance with "some" and "none" ~~were strongly~~ was correlated (Experiment 2). An individual differences study revealed a correlation between quantifier knowledge and implicature success (Experiment 3); a control study ruled out other factors (Experiment 4). Our findings suggest ~~that SI failures may in fact~~ some failures with scalar implicatures may be rooted in ~~a lack of quantifier~~ lack of semantic knowledge rather than general pragmatic or processing demands.

Keywords: scalar implicature, pragmatics, development

~~In speech, adult listeners routinely make inferences that go beyond the literal sense of an~~ Human language users have a remarkable ability: We are able to infer a speaker's intended meaning even when it was not explicitly conveyed by the literal meanings of the words in the utterance. For example, an adult who hears "I ate *some* of the cookies" would expect that the speaker did not eat *all* the cookies. Similarly, an adult ~~listener~~ who hears "I ate the sugar cookies," would most likely assume that the speaker ate *only* the sugar cookies, and not the other varieties. ~~These two statements use a~~ In both cases, listeners are inferring a speaker's intended meaning not from what was said, but from what was *unsaid*. More generally, pragmatic inferences like these are a critical part of language use (e.g. Clark, 1996; Levinson, 2000; Frank & Goodman, 2012).

Thus, to become fluent language users, children must not only learn the literal meanings of the words in an utterance, but also how to leverage these meanings to make inferences that go beyond the literal. In the two statements above, a speaker employs a weaker literal description ~~(in these cases, scalar and contextual, respectively)~~ to implicate that a stronger alternative is ~~true~~false. The first statement requires the listener to make a *scalar implicature*, which relies on lexical scales such as quantifiers ("some" vs. "all") or modals ("possibly" vs. "definitely") (Horn, 1972). The second statement requires an *ad-hoc implicature*, in which a stronger description is negated by a contextually weaker description~~; while~~ . In both instances, the listeners must reason not only about the semantics of the literal utterance, but also of the alternatives implicated in its pragmatic reading. A note about our use of the term "ad-hoc implicature": 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.

Both scalar and ad-hoc implicatures are abundant in language; it would be time-consuming and computationally costly to both a speaker and listener if all intended meaning had to be conveyed by a literal utterance. Thus, pragmatic listeners must be able to compute such implicatures readily. Making such inferences is a critical component of linguistic competence, permitting rich knowledge about intended meaning while minimizing speaker effort (Grice, 1975; Horn, 1984). How does such a crucial pragmatic skill develop, and what can it reveal more broadly about the developmental trajectory of the semantics–pragmatics interface (Papafragou & Musolino, 2003)?

Although much research has been devoted to the topic, when and how children acquire scalar implicatures remain unclear. While adults sometimes incur a processing cost in computing scalar implicatures along quantifier lexical scales like <SOME, ALL> (Bott & Noveck, 2004; Grodner, Klein, Carbary, & Tanenhaus, 2010; Huang & Snedeker, 2009a), by and large they compute implicatures reliably in a wide variety of situations. In contrast, early investigations showed that children's accuracy on these same scales was variable even until fairly late in development (Noveck, 2001). This result was intriguing because implicatures are an important case study for children's pragmatic reasoning more generally. The finding that even elementary-aged children struggled with certain types of pragmatic judgments suggested a surprising disconnect between pragmatics and other aspects of language development.

While this early work made an important contribution by identifying an area of difficulty for children, both the use of a truth-judgment task and the abstract, propositional nature of the materials (e.g., "Some giraffes have long necks") may have lead the paradigm to underestimate children's ability to make implicatures. In more recent investigations, children have shown a graded pattern of successes and failures across different tasks

~~(e.g., Guasti et al., 2005; Papafragou & Musolino, 2003; Papafragou & Tantalou, 2004)~~(e.g., Guasti et al., 2005; Papaf

Other evidence has suggested early competence in making some pragmatic judgments. For example, five-year-olds show evidence of recognizing that pragmatically-infelicitous statements deserve smaller rewards, when given a graded—rather than binary—judgment task (Katsos & Bishop, 2011). And some three- and four-year-olds are able to make ad-hoc (contextual) implicatures in a stripped–down referent-selection paradigm (Stiller, Goodman, & Frank, 2015). So although some methods still show notable failures ~~(e.g. Huang & Snedeker, 2009b)~~(e.g., Huang & Snedeker, 2009b), the evidence is more mixed as to what children's abilities are and precisely what causes the observed deficits.

Although much progress has been made in this area, one weakness of the current literature is its diversity. Table 1 provides a (non-exhaustive) summary of a number of influential experimental papers. First, a wide range of methods and measures—from referent selection to felicity judgment—have been used to assess children's implicature abilities. Second, most previous studies have targeted relatively wide age bins (12–18 months or wider), ranges that do not allow for precise developmental comparisons between studies. And finally, other paradigm-level differences might lead to widely varying levels of performance. For example, recent work indicates that the use of the partitive (e.g., "*Some* of the cookies") is a particularly strong cue for adults' scalar implicatures (Degen, 2015), but studies vary in their use of this ~~cue~~construction.

Thus, one goal of our current work is to consolidate insights from the previous literature by providing a direct comparison between ad-hoc and scalar implicatures in a simple referent selection task. This comparison allows for strong inferences about developmental differences between making scalar and ad-hoc implicatures. Additionally, a second goal of our current work

is to explore theoretical hypotheses about the source of children's surprising failures in scalar implicatures particularly. ~~(~~We follow previous work in suggesting that, if children succeed in ad-hoc implicatures and fail in scalar implicatures, it is very unlikely that a general pragmatic deficit explains previous findings. ~~)~~ The next section describes ~~possible candidate~~ other possible explanations.

Candidate Explanations for Failures in Scalar Implicature

One candidate explanation for children's failures to compute scalar implicature has emerged from previous literature. This idea, known as the *Alternatives Hypothesis*, suggests that children's performance may have less to do with general pragmatic knowledge per se, and more to do with their knowledge of the particular scales on which implicatures are computed. Barner and colleagues ~~(Barner & Bachrach, 2010; Barner, Brooks, & Bale, 2011)~~ (Barner & Bachrach, 2010; Barner et al., 2011) suggested 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~~ alternative words 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 provided a number of tests of this hypothesis. First, children's performance in implicature tasks increases when they have stronger access to ~~specifially~~ specifically lexical alternatives. Skordos and Papafragou (2016) made scalar alternatives salient within the experimental paradigm via a blocked design (e.g. "all" trials presented before "some" trials) and found that implicature performance increased. Second, even when lexical alternatives

are not presented directly, the use of more supportive paradigms or designs could be helpful. In particular, if referents corresponding to particular lexical alternatives are present on each trial, the presence of these referents could facilitate reasoning about alternatives. Supporting this idea, preschoolers showed some preliminary evidence of computing scalar implicatures for quantifiers in a forced-choice paradigm where different ~~quantification~~ quantificational scenarios were pictured (Miller, Schmitt, Chang, & Munn, 2005). And children were able to reason about ad-hoc scales both in a forced-choice paradigm (Stiller et al., 2015) and in a truth-value paradigm (Barner et al., 2011) when the relevant alternative objects were present in the trial context. Finally, recent work on inferences about disjunction argues that children struggle with the pragmatics of disjunction specifically due to failures in generating an adult-like set of alternatives, rather than general semantic or pragmatic deficits (Singh, Wexler, Astle-Rahim, Kamawar, & Fox, 2016; Tieu et al., 2016).

On the other hand, this body of research leaves open ~~an alternative~~ another possible explanation. Perhaps the problem is not with inferential alternatives generally, but instead a specific issue with quantificational alternatives. Perhaps children have trouble summoning to mind quantificational alternatives ~~because at varying developmental time points they~~ because—at varying times in development—they either 1) do not know the quantifiers, 2) do know them but cannot process them effectively, or 3) know them and can process them but have not grouped them into sets of lexical alternatives. In some sense, all three of these possibilities are consistent with the general spirit of the Alternatives Hypothesis, but problems 1 and 2 are perhaps more specific to quantifier meaning than otherwise supposed. Such quantifier-specific accounts might be most relevant for younger children. We return below to the idea that these accounts might not be mutually exclusive but might instead apply differently at different ages.

Supporting this general line of reasoning, quantifiers are a difficult lexical class for children to acquire. Unlike many other lexical classes (e.g., colors), quantifiers are not mutually exclusive with one another (and hence, they produce implicatures). In addition, quantifier semantics are relative to the total set size. For example, in a set size of 8, the quantifier "some" can be used felicitously to describe anywhere from 2 – 7 objects (Barner, Chow, & Yang, 2009) (and truth-functionally to describe sets 1 – 8), although adult-like responses tend to center around the mean of the set-size (Franke, 2014). These difficulties can be seen in an experiment by Hurewitz, Papafragou, Gleitman, and Gelman (2006), who found that 3–4-year-olds were only 75% correct in choosing the referent of "all" in a four-alternative forced choice. Thus, perhaps difficulties with the semantics of quantificational alternatives in particular are the problem, rather than alternatives more generally.

The Current Study

In ~~three~~ four experiments, we ~~explore~~ explored these ideas about the determinants of performance in implicature tasks using a novel paradigm. 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. This task gave children access to the visual alternatives in each trial (the three book covers) and the lexical alternatives across trials (either ad-hoc or scalar descriptions). Our design allowed us to counterbalance trial types (ad-hoc vs. scalar descriptions, crossed with implicature vs. unambiguous control targets) fully across participants, to examine both within-subject patterns of responses and between-subject developmental patterns, while also reducing the demands of the task.

In Experiment 1, we included both ad-hoc and scalar descriptions with implicature and control trials for each. Four–year-olds were at ceiling on ad-hoc trials (similar to previous work,

e.g. Stiller et al., 2015), but their performance on scalar implicature trials was very low.

In Experiment 2, we omitted ad-hoc trials. We found developmental increases in performance for each trial type, with higher performance on implicature trials for 4-year-olds in this scalar-only version. In both Experiments 1 and 2, we found an unexpected result: Children's pattern of responses on scalar implicature trials was bimodal and strongly correlated with their performance on "none" (scalar control) trials. This correlation suggested a general source of difficulty in comprehending these quantifiers beyond simply failing to make a scalar implicature.

In Experiment 3, we used an individual differences study to explore this correlation further. To ask whether quantifier knowledge was related to the specific task we used, we measured quantifier knowledge with a separate paradigm (Barner et al., 2009). We additionally assessed the alternative hypothesis that inhibitory control development, rather than quantifier alternative knowledge, might underlie the correlation between "none" and implicature performance; we found no evidence for this interpretation.

In Experiment 4, we tested the hypothesis that the salience of the *some–all* contrast might have been diminished by the inclusion of the strong alternative "none" in our implicature task. Contra this hypothesis—and consistent with previous work (Skordos & Papafragou, 2016), including this quantifier actually facilitated scalar implicature computation, rather than hindering it. Overall, our findings suggest that while preschoolers' computation of scalar implicatures can be supported by stronger recognition of the lexical alternatives, ~~their failures may be~~ some of their earlier failures are likely rooted in difficulty ~~comprehending and contrasting~~ understanding the relevant quantifiers.

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

Given the difficulty 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 both ad-hoc and scalar implicatures. In all scalar items, we used a partitive construction to increase the likelihood of making an implicature (Degen & Tanenhaus, 2015). All stimuli, data, and analyses for this and the subsequent experiments are available in a version-controlled, public repository at ()https://osf.io/mucf9.

Methods

Participants. Table 2 shows the demographic information for a planned sample of 48 children recruited from a university preschool. The preschool is an English language school , and children serving a high socioeconomic status and high educational achievement population. Children included in the sample were native speakers of English. Ethnicity information was not recorded for this sample. Children were recruited individually from the classroom, and tested in an individual testing room. Two children were excluded from the final sample for not completing the task, and one additional child was excluded due to experimenter error. No child completed more than one session of the task.

Stimuli. Experimental stimuli consisted of 18 sets of three printed pictures of book covers, each featuring four familiar items. 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 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 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. The experimenter emphasized that the child would only receive one clue about what book the experimenter was describing, and she had to use that clue to make her decision. All participants saw image sets (three books) 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. Table 3 shows the breakdown of trial types and sample scripts. Children did not receive feedback after the test trial.

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. At the start of the practice trial, the experimenter told the child "On the cover of my book, there's a TV." After children successfully completed the practice, they saw 18 test trials. At the beginning of every trial, the experimenter provided the child with either an ad-hoc or scalar description of one book, and instructed 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 the child should choose the single book she was describing.

In ad-hoc trials (eight total), the experimenter's descriptions of the target book used the

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 an 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."

Results

We found that all children performed at ceiling on ad-hoc implicature trials. In contrast to this success, however, they struggled significantly on scalar implicature ("some") and "none" trials. Children's accuracy on all trial types is plotted in Figure 2. On implicature trials, children's performance was coded as correct if they selected the image consistent with the implicature: the

single item (e.g., cats) on the ad-hoc trials, and the mixed item (e.g., cats and birds) on the scalar trials. Children were at ceiling making ad-hoc implicatures, which is consistent with previous research suggesting that children are able to succeed making such implicatures when they have access to the relevant lexical alternatives (Stiller et al., 2015). Children's performance across ad-hoc trials provides strong evidence that our novel paradigm is an appropriate measure for such items. In fact, the weak ad hoc utterance "there are cats" would be expected to produce less pragmatic effect than the stronger definite determiner version "the pictures are of cats." We selected the existential version for this study because we believed it was more felicitous. In contrast to their success in making ad-hoc implicatures, however, children ~~struggled~~ performed poorly on quantifier trials, ~~performing much worse for both~~ rarely choosing appropriately for either "some" ~~and~~ or "none" trials.

We ~~ran~~ fit 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. All mixed effects models were fit in R using the lme4 package. The model ~~specifications are~~ specification was as follows: (`correct ~ trial_type * condition * age + (trial_type | subject_ID)`). Age was centered for ease of model fit~~.~~, and the model included the maximal random effect structure consistent with our design, following the recommendations of Barr, Levy, Scheepers, and Tily (2013). Following our standard operating procedure, we began with all design-relevant fixed effects as random slopes and then iteratively removed coefficients until the model converged. We found that performance was significantly lower for scalar trials than ad-hoc trials ($\beta = -1.04$, $p = .001$), and that there was a significant interaction between condition and trial type, such that performance was significantly worse on scalar implicature trials ($\beta = -2.21$, ~~$p < .0001$). We~~ $p < .0001$). Unexpectedly, we also found a

significant 3-way interaction between condition, trial type, and age, such that performance on scalar implicature trials decreased with age (β = –4.2, p = .006). While this interaction may be unexpected based on children's tendency to improve on linguistic tasks with age, it is likely an artifact of our inclusion of both ad-hoc and scalar trials in the same task and so we do not interpret it further. We found no significant difference between age groups in an independent sample t-test on scalar implicature trials (t(46) = 1.28, p = .21), and no such interaction was found in Experiments 2 and 3. 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. There were no significant effects of controlling for within-trial item effects.

In a post-hoc ~~analysis~~ exploration of the data, we found ~~an unpredicted~~ a consistency in performance on "some" and "none" trials within subjects (Figure 3). ~~We~~ Although this pattern was surprising, we also observed this bimodal performance in a pilot sample (N = 23), so we had some reason to expect it in Experiment 1. To examine this pattern 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 result suggests children did not respond at chance in scalar trials, but instead were consistently either correct or incorrect. Additionally, children's success on "some" and "none" trials was correlated (r = .45, p = .001), such that children who performed better on "some" trials also tended to perform better on "none" trials. Performance on "none" and "all" trials (r = ~~.11~~.12, p = ~~.45~~.43) and "some" and "all" trials (r = .01, p = .95) was not correlated.

Discussion

The results of Experiment 1 indicated that children were easily able to make ad-hoc, but not scalar, implicatures in our task. 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 and lexical alternatives across the task, children still struggled to make scalar implicatures. In addition, performance did not increase across the course of the study, suggesting that the repetition of scalar alternatives across trials did not lead to greater levels of performance. ~~This~~

Our failure to find order effects differs from other work, where children only succeeded in generating scalar implicatures after rejecting an incorrect usage of "all" (Skordos & Papafragou, 2016). This ~~may stem from different processes children must undergo~~ difference could have resulted from age differences or task differences. The children in our study were by and large younger than those in Skordos and Papafragou (2016), so they may not have had strong enough quantifier representations to support the kind of priming posited in that earlier work. In addition, different processes may be at play when making a pragmatic inference in reference resolution (our task) versus evaluating the acceptability of a speaker's utterance (Skordos and Papafragou's task). In our task, children make an inference based on their own interpretation of the quantifier scale, while a truth-value judgment task, as in ~~(Skordos & Papafragou, 2016)~~Skordos and Papafragou (2016), requires that children accept or reject a statement based on a given true world state. Perhaps one task is more subject to priming than the other.

Even more ~~intriguing~~ surprising was the unexpected developmental change we observed on "none" trials. We included "none" as an unambiguous control quantifier, but found that children had lower performance 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). Further, this pattern is consistent with other work

indicating that children exhibit bimodal comprehension of this quantifier at least until 5.5 years of age (Barner et al., 2009).

~~It is possible that this pattern of performance stems from~~ One explanation for a number of our surprising findings is that they are a result of including both ad-hoc and scalar quantifier descriptions within ~~one~~ a single experimental session. Children's success on ad-hoc trials may have lead to a misinterpretation of scalar descriptions (e.g., "On the cover of my book, some of the pictures are cats") to ad-hoc descriptions ("On the cover of my book, there are some cats"). Presenting scalar descriptors in the same experimental session as other relevant and felicitous descriptions may alter scalar implicature comprehension, even in adults (cf. Degen & Tanenhaus, 2015). To explore the whether children's performance on scalar trials was 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

In Experiment 2, we pursued the possibility that children failed to make scalar implicatures as a result of competing ad-hoc descriptors in the same experimental session. Additionally, we expanded our target ~~age-range~~ age range to 3–5 years to more fully explore the developmental trajectory associated with making scalar implicatures.

Methods

Participants. Table 4 shows the demographic information for a new sample of 51 participants from the same university preschool. Of the children included in the sample, the majority were identified by caregivers as White (N = 17), multiracial (N = 7), or other (N = 21), with smaller proportions identified as Asian (N = 4) and Black (N = 2). One additional child was run but 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 3). 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 "some" (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 book 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. These three scripts were counterbalanced across participants.

Procedure. The procedure was identical to Experiment 1.

Results

In Experiment 2, children's performance in all trial types increased with age (Figure 4). Performance was highest in "all" trials across all age groups. However, performance was still significantly lower in both "none" and "some" trials (in comparison to "all, $p < .05$ for all tests); only 4.5–5-year-olds' performance for "none" was not significantly different than for "all" ($t(22) = 1.74$, $p = .11$.10). Children's performance in Experiment 2 was numerically, but 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), except for implicature performance in the oldest age group, which was significantly better ($t(34) = -2.54$, $p = .02$). All of these tests were relatively low in power, however, due to the small number of individuals in each bin.

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~~ effects that emerged were age and "some" trial types, such that performance increased across trials as children got older (β = ~~20~~3.5, p ~~<.001).~~ = .0002), but significantly decreased for implicature trials ($\beta = -3.3$, $p < .0001$.) 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 main effect was due to individual variability, such as in Experiment 1.~~

~~Consistent with the findings~~ Repeating the analyses of individual children's responses from Experiment 1, we ran Hartigan's dip test and again found significant bimodal ~~patterns of responses~~ responding for both "some" (D = .13, p <.0001) and "none" (D = .15, p <.0001) trials. ~~Once again,~~ Also, once again these trial types were highly correlated with one another across participants (r = .5, p = .0002).

~~Thus, as~~ As an exploratory analysis, we ran another version of the mixed effects model removing the random effect of trial type, as we hypothesized that our initial model did not find trial type effects due to the ~~correlational structure observed~~ correlation between trial types~~().~~. The specification of this model was `correct ~ trial type * age + (1 | subject)`. In addition to a main effect of age (β = 4.72, p <.0001), this model revealed a ~~conditional~~ condition effect: "some" trials (β = ~~–3.66~~-3.66, p <.0001) "none" trials (β = –3.1, p <.0001) were lower than "all" trials. 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 (β = ~~–2.84~~-2.84, p <.001), and "none" trials (β = ~~–1.95, p = .007~~-1.95, p = .007). 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 sum, while the effects we find with the models in Experiments 1 and 2 are generally consistent, estimates from the models are somewhat unstable due to individual differences.

In a further exploratory analysis, we investigated the particular ~~kinds of~~ errors that children made ~~in~~ on "some" and "none" trials (Figure 5). We found that children selected the ~~correct noun on both kinds of trials~~matching noun in both trial types, and chose the "all" option most frequently (where that noun named the only object pictured on the cover).

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. When presented with only scalar descriptions in Experiment 2, children's performance was numerically (although not significantly) better than in Experiment 1, with scalar implicature performance positively correlated with age. We still observed low performance in "some" and "none" trials. Additionally, we again found bimodal and correlated patterns of responses in these two trials, with children consistently failing or succeeding in both "some" and "none" trials. In sum, the results of Experiment 2 indicated that children struggle with making scalar implicatures beyond dealing with competing contextual descriptors. The cause of this failure ~~is~~ was not clear however, and individuals differed substantially.

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*). This idea is not consistent with classic pragmatic

theory (e.g., Horn, 1972), which posits that only alternatives that logically entail the current quantifier (e.g. “all” or “most”) take part in the implicature computation. Nevertheless, some recent work supports this possibility: Franke (2014) found in a model of pragmatic felicity that “none” was heavily weighted as an alternative in the scalar pragmatic computation for “some.” So there is some indirect evidence that children might need to know “none” to be able to make a scalar implicature with “some.”

In addition to understanding the extremes of the quantifier scale, another other possibility might also account for children’s failures. Children hearing “none of the pictures are cats” might simply match the word “cats” to the referent with the cats and fail to inhibit this match in favor of the correct alternative. This possibility is consistent with some work with adults on the comprehension of negation, where comprehenders have been posited to generate the positive match and then negate it (e.g. Kaup, Lüdtke, & Zwaan, 2006). We explored these two alternatives in Experiment 3, including measures of both inhibitory control and quantifier knowledge in the same session.

Experiment 3: Inhibitory control and quantifier knowledge measures

In an individual differences paradigm, we supplemented our implicature task with two additional tasks: an inhibitory control task, the Dimensional Change Card Sort (DCCS) (Zelazo, 2006); and a quantifier-knowledge task, Give-Quantifier (Barner et al., 2009). The DCCS 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). Our use of the DCCS as a metric of inhibitory control was motivated by the set-shifting inhibition required to succeed in DCCS. Unlike other executive function tasks (e.g., Go, No-Go) which

might tap into a more motoric type of inhibition, we hypothesize that DCCS success hinges on a participant's ability to inhibit a more salient response (a previously learned rule) by attending to the relevant linguistic information. Thus, our inclusion of the DCCS in Experiment 3 was driven by the similar attention to language required in both this task and our scalar implicature task.

In the Give-Quantifier task, a productive measure of quantifier knowledge, children give a quantity of items in response to a quantifier prompt. This task is well-suited to exploring children's grasp of quantifier semantics, allowing for free-response for both exact and inexact quantifiers (Barner et al., 2009). Thus, with these two tasks, we can assess the contributions of both quantifier knowledge and inhibitory control in driving children's ~~observed~~ performance in our scalar implicature task.

Methods

Participants. Table 5 shows the demographic information for a new planned sample of 72 children from the same university preschool; this sample was selected to have 80% power to detect correlations of $r > .3$. Of the children included in this sample, the majority were identified by caregivers as White (N = 27), multiracial (N = 17), and Asian (N = 18), with smaller proportions identified as Black (N = 4) and other (N = 6). Twelve additional children 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; these children were not excluded from analyses.

Stimuli. Stimuli for the implicature task were identical to Experiment 2. The materials for the DCCS, our inhibitory control measure, were drawn from the original methods paper (Zelazo, 2006). Fourteen laminated sorting cards (7 red rabbits and 7 blue boats) were put into two plastic sorting trays marked with either a target blue rabbit or red boat. 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~~ fruit (8 oranges, 8 bananas, and 8 strawberries) and a red plastic plate. Fruits were grouped together by kind at the start of each trial.

Procedure. Task order was counterbalanced across participants, and individual scripts for each task were also counterbalanced to avoid order effects. The tasks were done in a small room apart from the main classroom in individual sessions. The procedure for our implicature task was identical to Experiment 2. The experimenter asked ~~the child~~ children before the start of every task whether ~~she~~ they 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). Children were shown two plastic trays each marked with a target card (a blue rabbit and a red boat). At the beginning of the task, the experimenter explained that this was either the shape or color game, and that the cards had to be sorted according (e.g., in the color game, a red rabbit would be sorted into the red boat tray). After six trials, the experimenter told the participant that the rules had changed, and the cards had to be sorted by the other dimension (e.g., after the switch to the shape game, a red rabbit would be sorted into the blue rabbit tray).

In running the Give-Quantifier task, we followed the protocol of the original study (Barner et al., 2009), with the exception of limiting the quantifiers used in the task to "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 plate?"). 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; "some" trials were coded as correct if the child gave between 2 and 7 pieces, 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 6). We found significantly higher performance for all age groups with the quantifier "all" versus "some" and "none" in independent sample t-tests (p ~~<~~≤.02 for all tests), and again found developmental increases in performance for each quantifier. ~~Experiment 2 was not different than Experiment 3~~Performance for Experiment 3 was not significantly different than that observed in Experiment 2: Independent sample t-tests between age-groups for Experiments 2 and 3 generally did not yield ~~any~~ major 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~~(~~, $t(30) = 2.16$, p ~~<.05~~= .04).

As in Experiment 2, we ran Hartigan's dip test ~~in a post-hoc analysis and~~ on individual children's performance 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, we ran a partial correlation controlling for age, and found 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 due to an inhibitory control issue (i.e., making a response based solely on the target noun, regardless of quantifier used.) Overall, we found a developmental increase in DCCS performance from three to five years of age. Performance on post-switch trials was significantly

correlated with age (r = .28, p = .018), with 3–3.5 ~~-year-olds~~ and 3.5–4-year-olds at chance (3–3.5: $t(17)$ = -0.41, ~~t~~p ~~(~~= .68); 3.5–4: $t(17)$ = 1.69, p = .11), and only 4–5-year-olds performing significantly better than chance (4–4.5-year-olds: $t(17)$ = ~~3.05~~2.79, p ~~<.01~~= .013; 4.5–5-year-olds: $t(17)$ = ~~3.31~~3.61, p ~~<.01~~= .002). After controlling for age, we did not find a significant correlation between inhibitory control and performance on either "some" (r = .13, p = .26) or "none" trials (r = - .01, p = .93) in our implicature task.

Next, we turned our attention to the relation between children's performance on our scalar implicature task and quantifier knowledge. Children's performance on the Give-Quantifier task was very similar to performance on our implicature task, with all age groups performing at ceiling for the quantifier "all", and only older children succeeding on "~~most," "~~none," and "some" quantifier trials. Figure 9 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 through 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~~As in our scalar implicature task, we found that younger children in Give-Quantifier showed a bimodal and correlated pattern of response to the quantifier "none," with the majority of children giving either 0 or all 8 items in these trials, and gradually shifting to a more adult-like correct response by 4.5–5 years. Similar to performance on "some" trials in our scalar implicature task, we found that younger children gave all 8 objects in response to the prompt "some," and only the oldest age groups gave a more adult-like distribution of items in 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 partial-correlations controlling for age, we found that performance on Give-Quantifier and implicature tasks was ~~correlated~~related. 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" scalar implicature and "none" Give-Quantifier trials (r = .18, p = .14), although we did find a relation between performance on "none" scalar implicature and "some" Give-Quantifier trials (r = .35, p = .002).

In a further exploration of the relation between quantifier knowledge and scalar implicature performance, we examined both the particular kinds of errors that children made across both tasks, and how they were related. Figure 8 shows the breakdown of children's performance in our scalar implicature task on incorrect trials. As in Experiment 2, children exhibited evidence of making 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 9), in which younger children responded in a ~~bimodal fashion~~ largely bimodal manner for the items "some" and "none."

Discussion

In Experiment 3, we explored potential factors behind ~~children' s~~ 3–5-year-olds' difficulty making scalar implicatures, as well as with comprehension of the quantifier "none." 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 ~~relationamongst~~ relation 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 inhibitory control, at least at the level that we had statistical power to detect. We did find that children's performance on the Give-Quantifier and scalar implicature tasks were related, however, and children who failed "some" and "none" trials tended to do so across both tasks. This result indicates that quantifiers may be particularly difficult for children, even when removed from ~~the particular task we designed~~this particular scalar implicature task.

While it is clear that children in the age ranges we tested struggle with quantifier comprehension, the source of this developmental difficulty is less clear. Quantifiers are a difficult class of linguistic expression to acquire, due to their non-exact and varying corresponding set sizes (Hurewitz et al., 2006). Because children's definitions of these scalar terms are not solidified in the pre-school years, it is possible that that they are not able to use and contrast quantifiers on the <NONE − SOME − ALL> scale to succeed in ~~make a scalar implicature.~~ making scalar implicatures. That children should struggle with the semantics of quantifiers is not wholly unexpected; 5–year-old children and adults generate systematically different interpretations of sentences containing quantified noun phrases (Musolino, 1998). While children can be pushed towards adult-like interpretations of such sentences, their associated pragmatic readings of quantified noun-phrases are more tenuous than adults' (Musolino & Lidz, 2006). In addition to grappling with the semantics of quantifiers, children must also correctly interpret the scope involved in such utterances, which is particularly difficult for children when coupled with negation (Moscati & Crain, 2014; Musolino, 1998; Musolino & Lidz, 2006; Zhou & Crain, 2009). Thus, it seems likely that the pragmatic difficulties 3–5-year-olds face in interpreting utterances with quantifiers might be compounded by an inadequate grasp of their semantics.

It is possible, however, that the correlation we observe between children's performance on

our scalar implicature task and their quantifier knowledge might not be driven by incomplete or fragile quantifier knowledge, but rather, by appealing to a similar strategy across trial types. Recent work has shown that that 5–year-olds (slightly older than our sample) make scalar implicatures at adult-like levels when presented with either *some* and *all* or *some* and *none* when controlling for semantic knowledge of quantifiers (Skordos & Papafragou, 2014, 2016). While these results indicate that children seem to consider both "all" and "none" alternatives against which to compare the weaker quantifier "some," it is unclear whether this contrast remains as salient when considering all three quantifiers together. If children are able to compute implicatures along both the *some/all* and *some/none* scale, but fail to do so along *none/some/all*, it may point to the inclusion of both "none" and "all" mutually decreasing the contrast with "some."

In such a situation, children might choose to disregard the information provided by the quantifiers as under-informative, and thus resort to ignoring the quantifiers completely in all cases. They might choose to respond instead based solely on the noun phrase (NP) in each trial. Given the task similarities between our scalar implicature task and Give-Quantifier, it is possible that children who are recruiting this strategy for one task might also do so for the other.

The use of this common strategy for both tasks would naturally lead to a strong correlation between scalar implicature performance and quantifier knowledge, but not between either task and the DCCS. While we believe that the DCCS requires children to employ a similar strategy in both our scalar implicature and Give-Quantifier tasks (namely, inhibiting a response by explicitly attending to the pertinent linguistic information), the two tasks differ in that children only have linguistically available to them the relevant dimension along which to sort the card in DCCS. Thus, children could not apply the NP-matching strategy here.

The question, therefore, is whether the full set of alternatives in our scalar implicature task

(*none/some/all*) in fact obscures the relevant contrast between the alternatives. If this is the case, then children's poor performance in our task might not be reflective of their pragmatic abilities, but rather the result of a task-specific NP-matching strategy. To explore this possibility, we ran an additional control study on our scalar implicature task including only "some" and "all" as lexical alternatives.

Experiment 4

To explore whether children's lower performance on our scalar implicature task was driven by the inclusion of two relevant alternatives ("all" and "none"), we conducted a separate control study including only one relevant lexical alternative ("all") to increase the salience of its contrast to the weaker "some." Given our previous results showing generally poor performance on this task with younger children, and other work showing stronger scalar implicature computation with older children (Skordos & Papafragou, 2014, 2016), we included only 4–5-year-olds in our sample.

Methods

Participants. Table 6 shows the demographic information for a new sample of 38 participants recruited from a local children's museum. A smaller sample of 36 was planned but more children than expected were recruited. This museum draws from a diverse community in a large city; our previous work suggests that museum-goers tend to skew towards having high educational attainment. Of the children included in the sample, the majority were identified by caregivers as White (N = 17), asian (N = 12), multiracial (N = 5) or other (N = 4). Twenty-four additional children were run but not included in analyses due to low English exposure (<75%), age outside of the planned range, or parental interference. Two children in our sample have one trial each excluded from analysis due to experimenter error. Testing took place in a small room

away from the main museum floor.

Stimuli. Stimuli were identical to those in the scalar implicature task used in Experiments 1–3. The only modifications made were reducing the number of total trials to 12, with six trials per quantifier ("some" and "all"). Once again, we randomly assigned participants to one of three lists, which pseudo-randomized quantifier order and controlled for position of the correct book referent. To control for correct referent location, we repositioned several book covers in the original stimulus set.

Procedure. The procedure was identical to Experiments 1–3.

Results

When presented with only "some" and "all" as alternatives in this task, 4–5-year-olds performed significantly worse on implicature trials than "all" trials ($t(74) = -15.972$, $p < .0001$). An independent t-test did not reveal a significant difference for performance between age groups for either "all" ($t(36) = .5$, $p = .62$) or "some" ($t(36) = -0.23$, $p = .82$). Surprisingly, both age groups performed significantly worse on implicature trials in comparison to Experiment 3 (4–4.5-year-olds: $t(34) = 2.19$, $p = .035$; 4.5–5-year-olds: $t(36) = 3.7$, $p = .0007$). Figure 10) shows performance for each trial type and age group. When making an incorrect selection on "some" trials, children overwhelmingly chose the book consistent with "all" (N = 186 trials, out of a total 190 incorrect), while choosing the book consistent with "none" only once.

We next fit a logistic mixed effects model predicting a correct response as an interaction of age and prompt. The model specification was `correct ~ trial type * age + (trial type | subject)`. Age was centered for ease of interpretation. We found a significant negative effect of trial type, such that performance was significantly worse on "some" trials in comparison to

"all" (β = -11.19, p <.0001). There was no significant interaction between age and trial type.

Discussion

In Experiment 4, we tested the hypothesis that the inclusion of "none" in our referent-selection task decreased the saliency of the contrast between "some" and "all," prompting children to disregard underinformative quantifier information, and resort instead to NP-matching strategy. Surprisingly, we found that excluding "none" as a lexical alternative in this task actually *reduced* performance on implicature trials.

This finding is interesting for several reasons. First, it provides further support for "none" being under consideration as an alternative when drawing a scalar implicature from "some" (cf. Skordos & Papafragou, 2016). When children are familiar with the semantics of "none," and when it is available as an alternative, children benefit from its inclusion in this task. Second, these results suggest that the inclusion of another alternative might help "unstick" children from their bias towards choosing the referent consistent with "all" in our task. Some recent work using the same task as our study has suggested that children might have to overcome such a bias in making scalar implicatures (Schneider & Frank, 2016).

Overall, we did not find evidence to suggest that the inclusion of "none" in our task hindered children's performance in making scalar implicatures. Rather, we found quite compelling evidence that its inclusion actually *helps* children in our task, and may overtly draw their attention to an alternative that should be under consideration during this computation.

General Discussion

Learning to infer what was said by considering what might have been said—considering contextual alternatives in pragmatic inference—is a key part of being a fluent and proficient

language user. Implicatures like those we studied here are an important case study of such pragmatic inferences. Implicatures are frequent in adult speech, reducing speaker effort while maximizing listener comprehension (Grice, 1975; Horn, 1984). Why do children continue struggle to compute scalar implicatures until relatively late in language development despite their success in other pragmatic domains?

We designed a simple task to investigate patterns of pragmatic development involved in making scalar implicatures 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 (Stiller et al., 2015) ~~, though we~~ and found poor performance on scalar implicatures even using a simplified paradigm. In Experiment 2, we removed competing ad-hoc descriptions from our implicature task, and found marginal increases in preschoolers' comprehension of all scalar quantifier terms. We still observed correlated difficulty with the quantifiers "some" and "none," however. In Experiment 3, we explored two possible explanations for this pattern of performance, inhibitory control and quantifier knowledge, and found ~~evidence that~~ a significant relation between children's ~~difficulties on~~ knowledge of quantifiers and their performance in our implicature task~~are rooted in a lack of quantifier knowledge~~, but not between inhibitory control and scalar implicature performance. In Experiment 4, we found that the inclusion of "none" as a lexical alternative in our task significantly improves, rather than ~~inhibitory control. Our findings suggest that while 4-year-olds are able to compute scalar implicatures with support from contextual cues, their performance is fragile~~ inhibits, scalar implicature computation. Taken together, our work suggests that at least some of the difficulties younger children encounter when making scalar implicatures lie in their fragile grasp of the quantifiers involved.

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 increase the precision of our measurements 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 found evidence for a link between poor quantifier knowledge and the ability to make scalar implicatures.

The Alternatives Hypothesis

Our findings provide some support for the Alternatives Hypothesis (Barner & Bachrach, 2010; Barner et al., 2011). ~~In particular~~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. ~~In addition~~Second, the presence of ad-hoc trials decreased scalar ~~peformance~~ performance in Experiment 1 compared with Experiment 2, suggesting that children might have recognized the ad-hoc descriptions as alternatives that competed with the quantifiers. Third, the presence of the putative alternative "none" in the task appeared to increase performance (in the comparison between Experiments 3

and 4).

On the other hand, another 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~~the experiments. We had expected that, if children's difficulties with scalar implicature were due to a lack of recognition of the contrastive relation between "some" and "all," that this relation would be revealed by the two words' consistent use in contrasting references over the course of the many trials that each child completed (cf. Skordos & Papafragou, 2016). 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 other inferential issues. One possible reconciliation of these results is that the younger children in our task were not as sensitive to priming by repeated presentation of alternatives.

~~Further, the correlated responses for "some" and "none" trials in both Experiments 1 and 2 suggested that children's difficulty with these quantifiers could have a common root, either in quantifier knowledge or inhibitory control. Our data in Experiment 3 provided evidence against a purely inhibitory account, leaving the quantifier knowledge hypothesis as one possible contender.~~

Lexical Scales

Although "none" is not typically considered part of the same Horn scale as "some" and "all," it is nonetheless a salient member of the same lexical class. Recent work has suggested that "none" is in fact a salient inferential alternative, at least in some models of implicature (Franke, 2014). Additionally, recent developmental work has indicated that, when controlling for quantifier semantic knowledge, 5–year–old children make scalar implicatures with "some" at adult-like

levels when it is contrasted with "none" (Skordos & Papafragou, 2014, 2016). This result is further underscored by younger children's lower performance in Experiment 4 when "none" was not an available lexical alternative.

For children who are familiar with these quantifiers, "none" appears to be important and relevant alternative to "some" that is compared along the same lexical scale when drawing an implicature. Thus, perhaps the strong correlation we observed between "some" and "none" is in fact due to the ~~lack of a good semantics~~ incomplete semantic knowledge for "none" leading to failures. This would not be wholly unexpected; children do not demonstrate complete knowledge of "none" until fairly late in development (Barner et al., 2009). In other words, if children either don't know or can't process all the quantifiers in the lexical class, they may fail to reason appropriately about implicatures.

~~Our results suggest that children's difficulties computing scalar implicatures lie in incomplete quantifier knowledge, and that children who failed to make implicatures similarly struggled in comprehending the quantifier scale.~~

Quantifier Knowledge

While our findings are generally consistent with the Alternatives Hypothesis, ~~we~~ our results suggest that quantifier knowledge specifically is a bottleneck which constrains younger children's ability to recognize and contrast relevant ~~lexical alternatives in generating implicatures.~~

~~Several limitations in the current study provide directions for further work. First, we did not record~~ alternatives along a lexical scale. Based on the results from this work, incomplete quantifier knowledge is at least one factor driving children's ~~response time on the scalar implicature task. Perhaps a response latency measure might yield more information about the particular processing and inhibitory demands of scalar implicatures. Additionally, our quantifier~~

~~knowledge measure (Barner et al., 2009) may present~~ struggles with scalar implicatures. Quantifiers as a lexical class are surprisingly difficult to acquire (cf Hurewitz et al., 2006), and it is unsurprising that preschoolers grapple with their meanings until fairly late in development.

Children's troubles with scalar implicatures are not confined solely to quantifiers, however. Developmental work on logical connectives indicate that children also have difficulties computing a scalar implicature with *or* to generate a disjunctive, rather than conjunctive interpretation (Singh et al., 2016; Tieu et al., 2016; Zhou, Romoli, & Crain, 2013). Interpreting a statement such as "The girl has a balloon or a ball" as disjunctive requires the listener to draw a scalar implicature that the girl has either a balloon or a ball, but does not have both (Singh et al., 2016). Children favor inclusive over adult-like exclusive readings of *or* in these tasks, not because they don't understand the semantics of *or* in this task, or are generally unable to compute implicatures, but because they cannot use the full set of alternatives to draw the implicature (Singh et al., 2016; Tieu et al., 2016; Zhou et al., 2013).

These difficulties with disjunction further highlight the difficulties of accessing alternatives to generate pragmatic inferences even when children understand their semantics. Thus, while we suggest that quantifier knowledge is an important factor limiting children's success in computing scalar implicatures, we also acknowledge that once children have full access to their semantics, there may still be difficulties in consistently generating these alternatives to make pragmatic inferences.

Limitations

Our studies have a number of limitations that point the way towards future work. First, it is difficult to measure quantifier semantics independently from pragmatics. In fact, the quantifier knowledge measure we used (from Barner et al., 2009) may have presented many of the same

pragmatic hurdles as our implicature task. ~~For example, the~~ The syntactic construction of prompts in ~~Give-Quantifier are very similar~~the two tasks were very similar, for example (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."). ~~It is also possible that, similarly to adults, children's scalar implicatureperformance may also be affected by non-scalar alternatives, such as "one" or "two" (Degen & Tanenhaus, 2015). Indeed, children show some evidence of comprehending quantifiers like~~ This similarity may have prompted children with low quantifier knowledge to rely on a similar NP-matching strategy in both tasks, leading to the observed relationship. Although we tried to test one prediction of the NP-match account in Experiment 4, we cannot completely rule it out.

Second, we did not compare the younger children in our study to an older population who might have different difficulties with scalar implicature. In other work, however, we used our referent-selection paradigm with 5–6.5-year-olds, and again found low performance with "some" and "none" ~~when presented in the same truth-value judgment task as exact number terms (Barner et al., 2009). Future work could explore other more distinct measures of quantifier semantics, as well as the role of other alternatives involved in making scalar implicatures~~(Schneider & Frank, 2016).

Finally, as with nearly all work in this tradition, our studies here have focused on quantifier comprehension in relatively high socioeconomic status populations who are likely exposed to a wide variety of rich, quantified language in their environment. More work is needed to understand the generality of the developmental account we proposed here.

Conclusion

In sum, our work suggests that 4–5-year-old 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 trouble with quantifiers is not confined to making scalar implicatures, but extends to~~ during this period. We suggest that, along with trouble contrasting the semantics of lexical alternatives, one additional cause of these difficulties is children's ~~knowledge of the quantifier "none" as well~~incomplete knowledge of quantifier semantics.

~~Tables and Figures~~

References

Barner, D., & Bachrach, A. (2010). Inference and exact numerical representation in early language development. *Cognitive Psychology*, *60*(1), 40 - 62. doi: 10.1016/j.cogpsych.2009.06.002

Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition*, *118*(1), 84 – 93. doi: 10.1016/j.cognition.2010.10.010

Barner, D., Chow, K., & Yang, S.-J. (2009). *Cognitive psychology*, *58*(2), 195–219. doi: 10.1016/j.cogpsych.2008.07.001

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*(3), 255–278. doi: doi:10.1016/j.jml.2012.11.001

Bott, L., & Noveck, I. A. (2004, 10). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language*, *51*(3), 437–457. doi: 10.1016/j.jml.2004.05.006

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

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. doi: 10.3765/sp.8.11

Degen, J., & Tanenhaus, M. K. (2015). Processing scalar implicature: A constraint-based approach. *Cognitive science*, *39*(4), 667–710. doi: 10.1111/cogs.12171

Frank, M., & Goodman, N. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998.

Franke, M. (2014). Typical use of quantifiers: A probabilistic speaker model. In *Proceedings of the 36th annual conference of the cognitive science society* (pp. 487–492).

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

Grodner, D., Klein, N., Carbary, K., & Tanenhaus, M. (2010). *Cognition*, *116*(1), 42–55. doi: 10.1016/j.cognition.2010.03.014

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*(5), 667. doi: 10.1080/01690960444000250

Horn, L. (1972). *On the semantic properties of logical operators* (Unpublished doctoral dissertation). UCLA.

Horn, L. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. *Meaning, form, and use in context*, *42*.

Huang, Y. T., & Snedeker, J. (2009a). Online interpretation of scalar quantifiers: Insight into the semantics/pragmatics interface. *Cognitive Psychology*, *58*(3), 376 - 415. doi: 10.1016/j.cogpsych.2008.09.001

Huang, Y. T., & Snedeker, J. (2009b). Semantic meaning and pragmatic interpretation in 5-year-olds: Evidence from real-time spoken language comprehension. *Developmental Psychology*, *45*(6), 1723-1729. doi: 10.1037/a0016704

Hurewitz, F., Papafragou, A., Gleitman, L., & Gelman, R. (2006). Asymmetries in the acquisition of numbers and quantifiers. *Language Learning and Development*, *2*(2), 77-96. doi: 10.1207/s15473341lld0202˙1

Katsos, N., & Bishop, D. (2011). Pragmatic tolerance: Implications for the acquisition of

informativeness and implicature. *Cognition*, *120*(1), 67-81. doi: 10.1016/j.cognition.2011.02.015

Kaup, B., Lüdtke, J., & Zwaan, R. A. (2006). Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? *Journal of Pragmatics*, *38*(7), 1033-1050. doi: 10.1016/j.pragma.2005.09.012

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

Miller, K., Schmitt, C., Chang, H., & Munn, A. (2005). *Young children understand some implicatures.* Proceedings of the 29th Annual Boston University Conference on Language Development: Cascadilla Press.

Moscati, V., & Crain, S. (2014). When negation and epistemic modality combine: The role of information strength in child language. *Language Learning and Development*, *10*(4), 345-380.

Musolino, J. (1998). *Universal grammar and the acquisition of semantic knowledge: An experimental investigation into the acquisition of quantifier-negation interaction in english* (Unpublished doctoral dissertation). University of Maryland.

Musolino, J., & Lidz, J. (2006). Why children aren't universally successful with quantification. *Linguistics*, *44*(4), 817-852.

Nordmeyer, A., & Frank, M. (2014). The role of context in young children's comprehension of negation. *Journal of Memory and Language*, *77*, 25-39. doi: 10.1016/j.jml.2014.08.002

Noveck, I. (2001). When children are more logical than adults: experimental investigations of scalar implicature. *Cognition*, *78*(2), 165-188. doi: 10.1016/S0010-0277(00)00114-1

Papafragou, A., & Musolino, J. (2003). Scalar implicatures: experiments at the

semantics-pragmatics interface. *Cognition*, *86*(3), 253-282. doi: 10.1016/S0010-0277(02)00179-8

Papafragou, A., & Tantalou, N. (2004). Children's computation of implicatures. *Language Acquisition*, *12*, 71-82. doi: 10.1207/s15327817la1201˙3

Schneider, R. M., & Frank, M. C. (2016). A speed-accuracy trade-off in children's processing of scalar implicatures. *Proceedings of the 38th Annual Meeting of the Cognitive Sciences Society*.

Singh, R., Wexler, K., Astle-Rahim, A., Kamawar, D., & Fox, D. (2016). Children interpret disjunction as conjunction: Consequences for theories of implicature and child development. *Natural Language Semantics*, *24*(4), 305-352.

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.

Skordos, D., & Papafragou, A. (2016). Children's derivation of scalar implicatures: Alternatives and relevance. *Cognition*(153), 6-18. doi: 10.1016/j.cognition.2016.04.006

Stiller, A., Goodman, N., & Frank, M. (2015). Ad-hoc implicature in preschool children. *Language Learning and Development*, *11*(2), 176-190. doi: 10.1080/15475441.2014.927328

Tieu, L., Yatushiro, K., Cremers, A., Romoli, J., Sauerland, U., & Chemla, E. (2016). On the role of alternatives in the acquisition of simple and complex disjunctions in french and japanese. *Journal of Semantics*, *33*(3).

Zelazo, P. D. (2006). The dimensional change card sort (dccs): a method of assessing executive function in children. *Nature Protocols*, *1*, 297-301. doi: 10.1038/nprot.2006.46

Zhou, P., & Crain, S. (2009). Scope assignment in child language: Evidence from the acquisition

of chinese. *Lingua*, *119*(7), 973-988.

Zhou, P., Romoli, J., & Crain, S. (2013). Children's knowledge of free choice inferences. *Semantics and Linguistic Theory*, *23*, 632-651.

Study	Scale(s)	Ages	Measure	Sentence Type	Main Finding
Noveck (2001)	non-necessity–necessity, possibility–impossibility, some–all	5;1–5;11, 7;1–8;0, 9;0–9;5, 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, two–three, start–finish	4;11–5;11 (Study 1) 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, ad-hoc, 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~~ compute 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) 3;6–5;10 (Study 2)	Direct Instruction Task (Study 1); Picture Matching Task (Study 2)	"Make some faces HAPPY/Make SOME faces happy/Make some HAPPY faces" (Study 1), "Show 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, two–three	5;2–6;1 (Study 1) 5;5–6;9 (Studies 2 & 3)	Eye-tracking 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, 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, 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~~some" if they first heard "all" falsely refer to quantity (only 3/4 ~~blackouts~~ blickets had crayons), but not ~~is~~ if "all" referred falsely to the objects (e.g., "all of the ~~blackest~~ blickets have a scarf").

Table 1
Review of previous literature on children's comprehension of implicatures.

Age group	N	N Female	N Male	Mean	Median	SD
4.0 – 4.5-year-olds	24	12	12	4.2	4.19	0.14
4.5 – 5.0-year-olds	24	17	7	4.74	4.73	0.16

Table 2
Participant age information for Experiment 1.

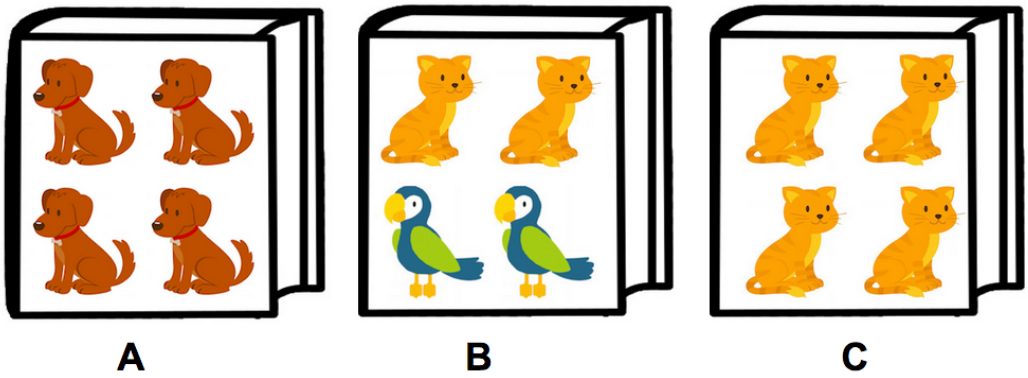

Figure 1. Example trial stimuli used in all experiments. Children received a clue from the experimenter about which book she had in mind and responded based solely on the clue; this was either an ad-hoc or a scalar description of a book with either an unambiguous or implicature target.

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 3
Study ~~designs~~ design for ~~Experiments 1~~our scalar implicature task, ~~2, and 3,~~ using script examples for the stimulus set pictured in Figure 1.

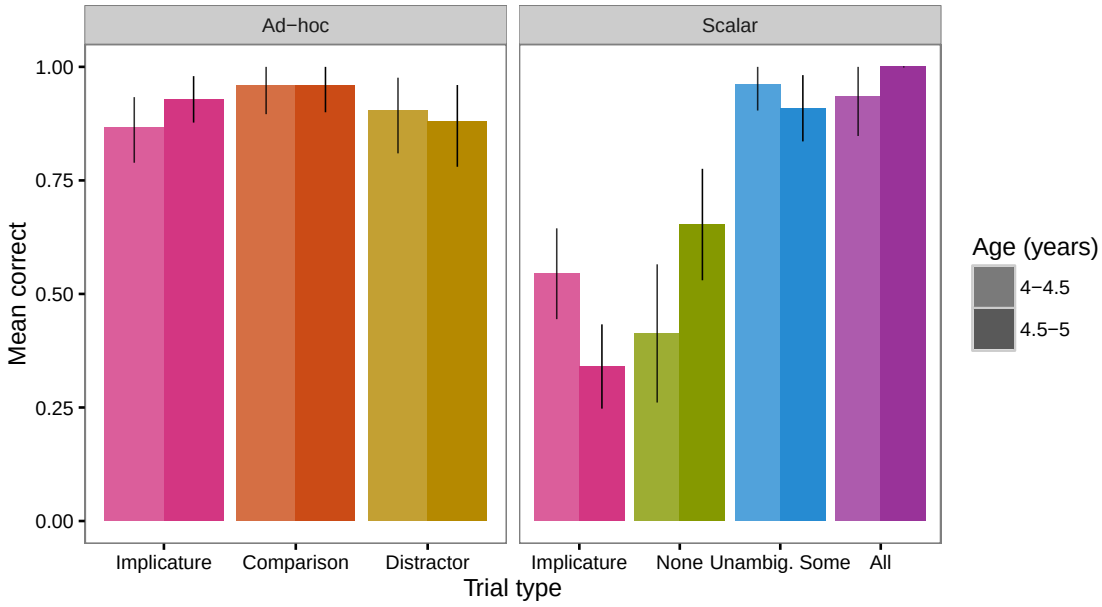

Figure 2. Proportion of correct responses by each age group across all trial types and split by implicature type ~~.~~ for Experiment 1. Error bars show 95% confidence intervals computed by non-parametric bootstrap.

Age group	N	N Female	N Male	Mean	Median	SD
3.0 – 3.5-year-olds	12	6	6	3.4	3.35	0.1
3.5 – 4.0-year-olds	13	6	7	3.8	3.67	0.12
4.0 – 4.5-year-olds	14	9	5	4.3	4.24	0.1
4.5 – 5.0-year-olds	12	3	9	4.8	4.63	0.15

Table 4
Participant ~~age~~ information for Experiment 2.

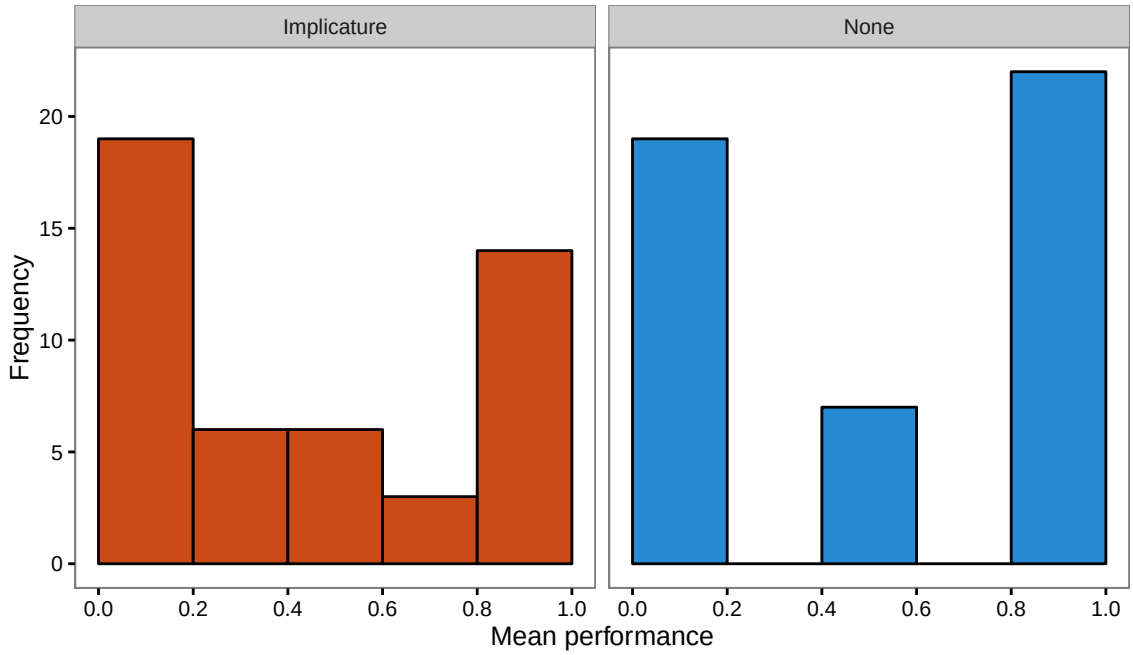

Figure 3. Frequency of mean performance in Experiment 1, split by condition.

Age group	N	N Female	N Male	Mean	Median	SD
3.0 – 3.5-year-olds	18	12	6	3.23	3.24	0.13
3.5 – 4.0-year-olds	18	10	8	3.73	3.8	0.18
4.0 – 4.5-year-olds	18	7	11	4.16	4.11	0.11
4.5 – 5.0-year-olds	18	9	9	4.7	4.68	0.14

Table 5
Participant ~~age~~ information for Experiment 3.

Age group	N	N Female	N Male	Mean	Median	SD
4.0 – 4.5-year-olds	18	14	4	4.22	4.21	0.14
4.5 – 5.0-year-olds	20	14	6	4.79	4.82	0.17

Table 6
Participant information for Experiment 4.

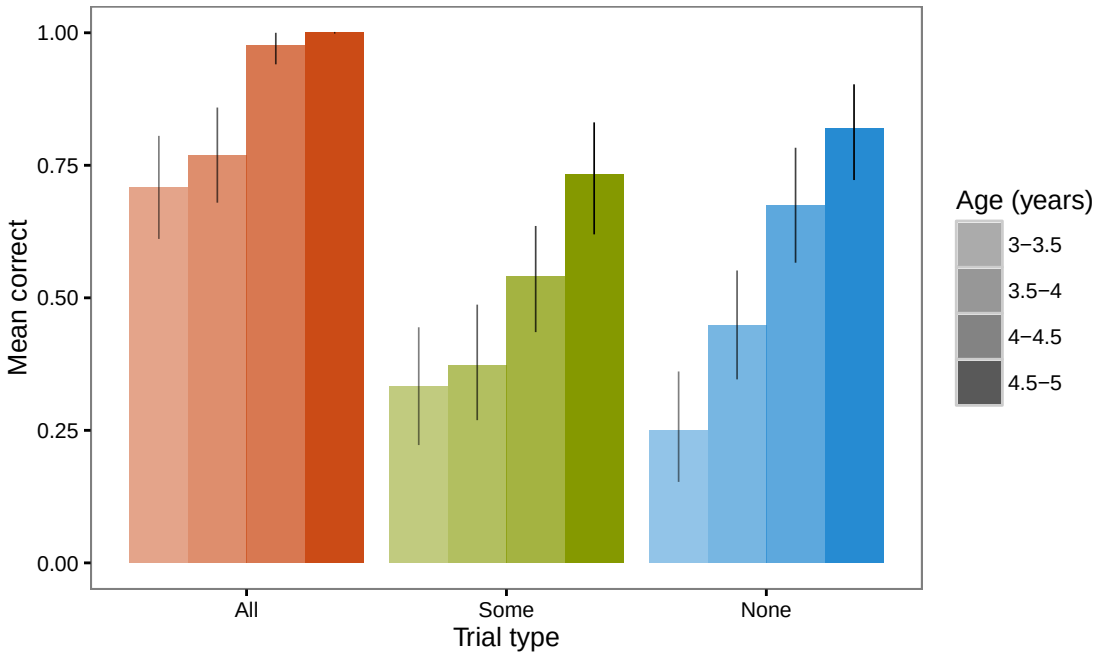

Figure 4. Proportion of correct responses by each age group across all scalar trial types ~~.~~ for Experiment 2. Error bars show 95% confidence intervals computed by non-parametric bootstrap.

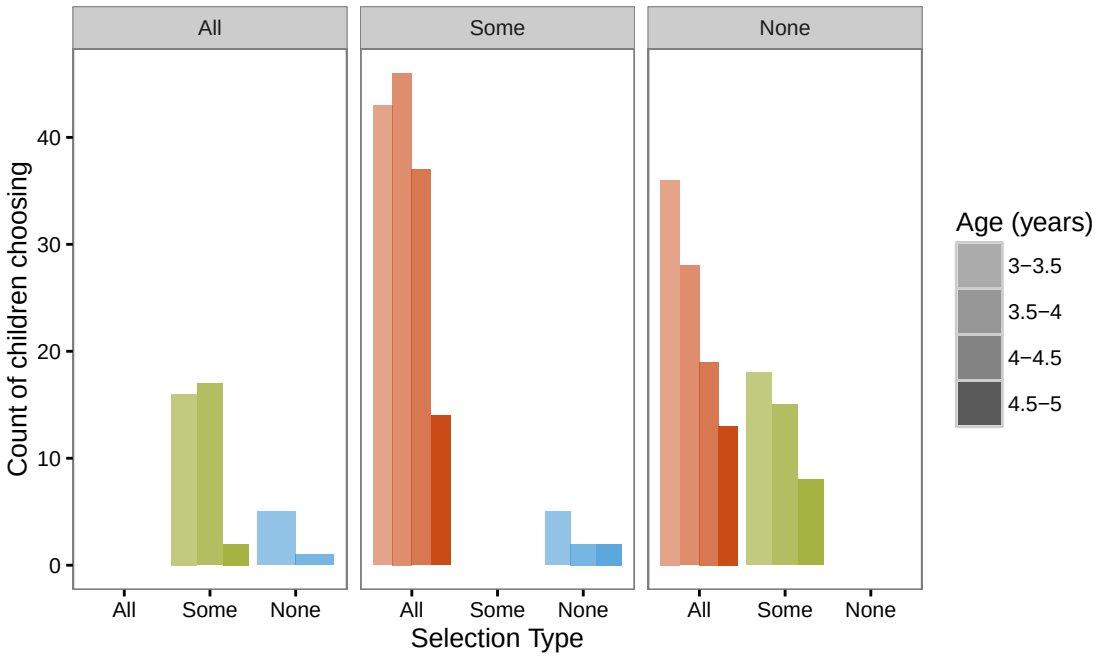

Figure 5. Scalar implicature error analysis ~~.~~ for Experiment 2. Count of children choosing an alternative on incorrect trials, faceted by trial type ("all," "some," and "none") and split by age group.

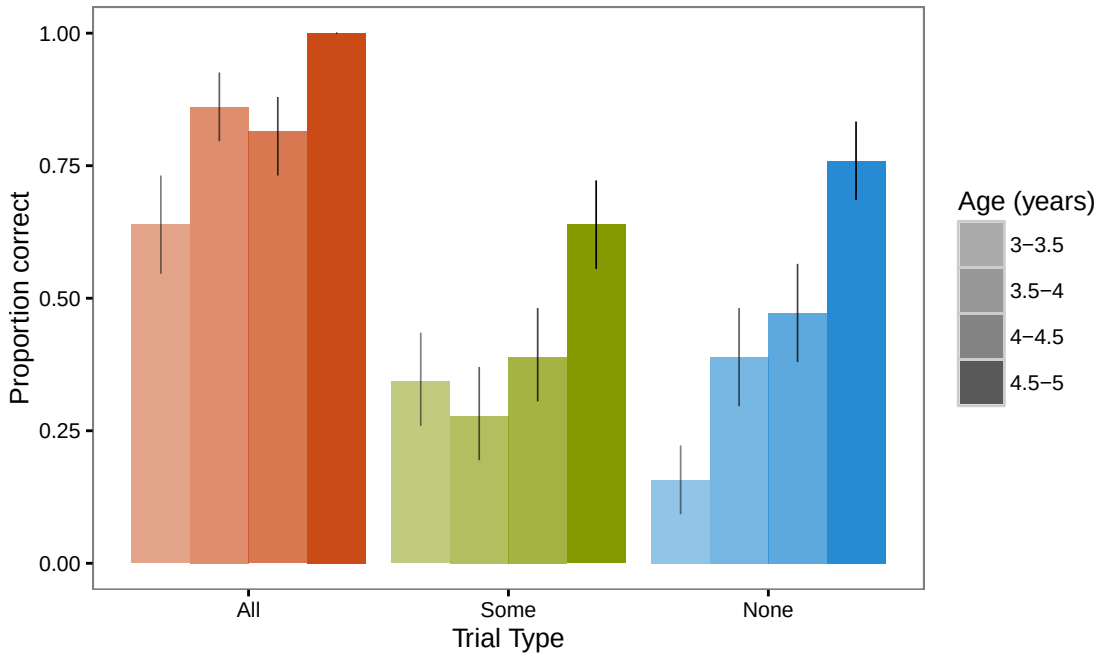

Figure 6. Proportion of correct responses by each age group across all trial types for the scalar implicature task ~~.~~ in Experiment 3. Error bars show 95% confidence intervals computed by non-parametric bootstrap.

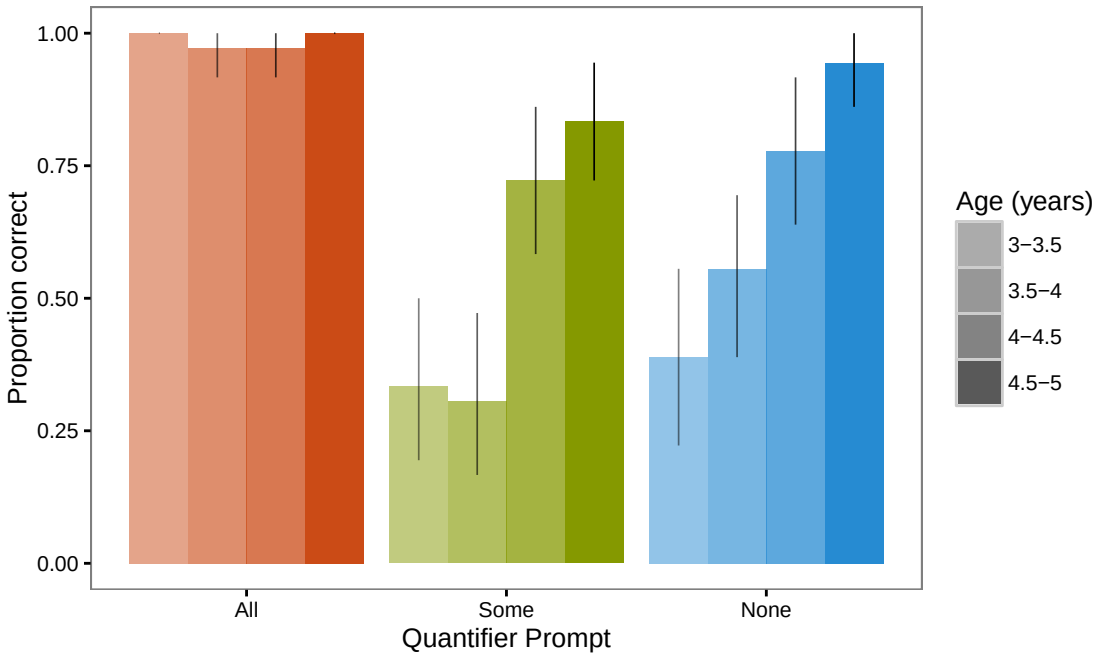

Figure 7. Proportion of correct responses by each age group for the Give Quantifier task in Experiment 3, ~~with~~ for quantifier prompts *all, most, none* and *some*. Error bars show 95% confidence intervals computed by non-parametric bootstrap.

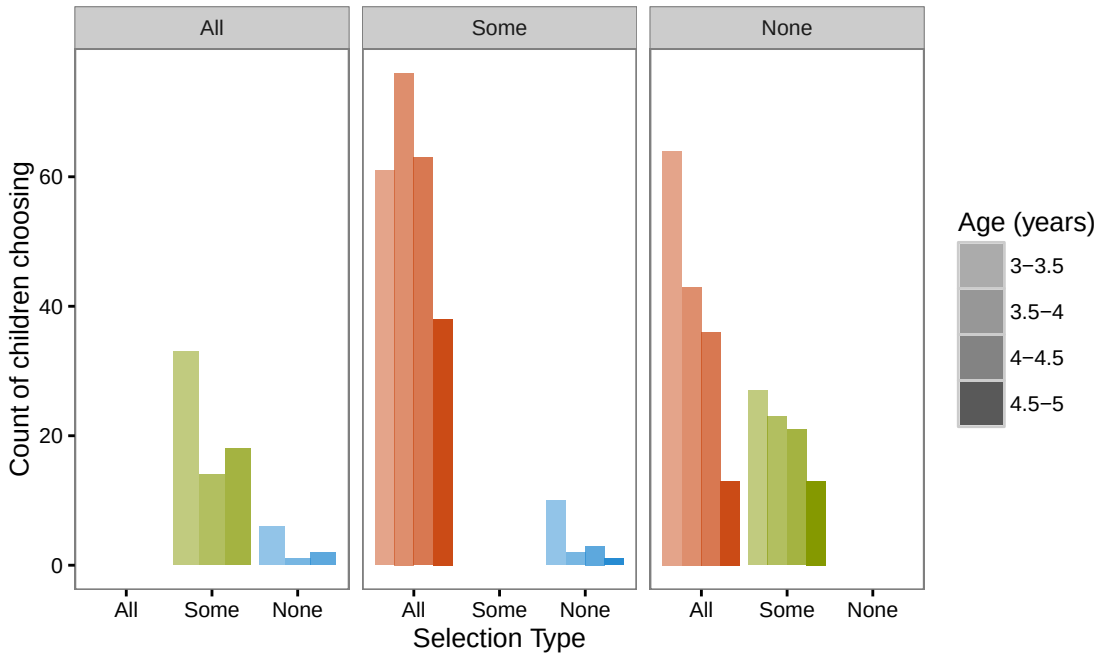

Figure 8. Scalar implicature error analysis ~~.~~ for Experiment 3. Count of children choosing an alternative on incorrect trials, faceted by trial type ("all," "some," and "none") and split by age group.

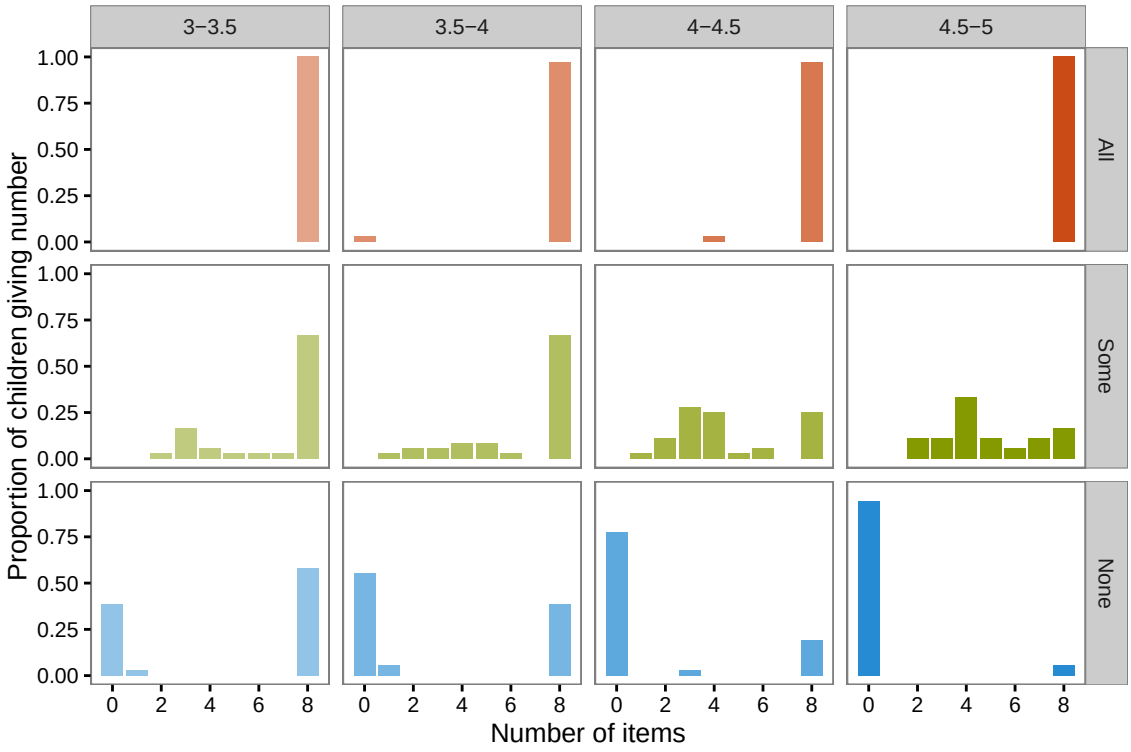

Figure 9. Give-Quantifier performance ~~.~~ for Experiment 3. Proportion of children giving numbers of items faceted by age group and quantifier prompts (“all,” “none,” and “some”).

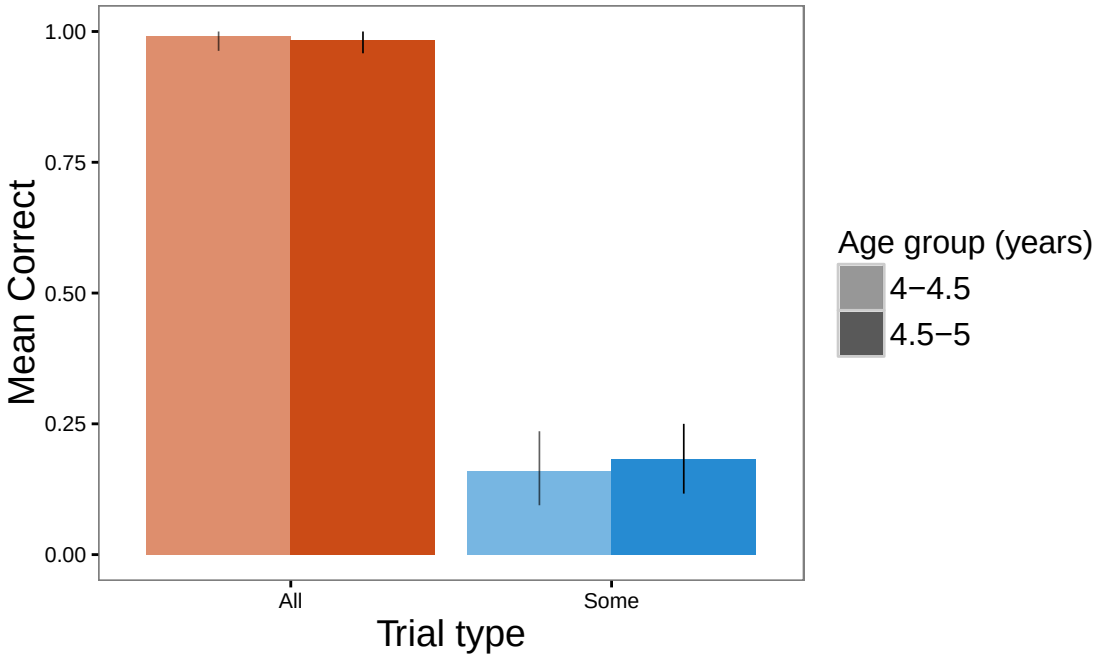

Figure 10. Scalar implicature performance in Experiment 4. Plotting conventions are as above.