Sources of developmental change in pragmatic inferences about scalar terms

Alexandra C. Horowitz
ahorowit@stanford.edu
Department of Psychology
Stanford University

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Abstract

Pragmatic implicatures—inferences that a weak statement (e.g. "some of my siblings") implies that a stronger one ("all of my siblings") could not be used—are a popular case study of children's pragmatic development. A growing literature suggests that children can make implicatures under some (but not all) conditions, but their performance varies widely across tasks, and few datasets allow direct comparisons between implicature types. To address this issue, we combine different types of implicatures and control trials into a single, simple paradigm. In Experiment 1, we included both ad-hoc (contextual) and scalar (quantifier) descriptions, and found that 4-year-olds were at ceiling in ad-hoc trials but had difficulty with scalar implicatures using quantifiers. In Experiment 2, 4-year-olds' performance increased when we included only scalar trials, but was still low. Across both datasets, we found a positive correlation between performance for "some" and "none" quantifiers. Our work provides more precise developmental data on the emergence of different implicature computations and illustrates that preschoolers' recognition of implicatures relates both to their comprehension of particular lexical items and also their recognition of relevant alternatives.

Keywords: Pragmatics; implicature; language development.

Introduction

Speakers tend to produce utterances that are informative given their intended meaning. Implicatures are instances in which a weak description (e.g. "I did some of my homework") implies that a stronger alternative (that I did *all* of it) is not true, or else a cooperative communicator would have used the stronger case. *Scalar* implicatures rely on lexical scales, or sets of related terms that are graded in meaning such as quantifiers ("some" vs. "all"), modals ("possibly" vs. "definitely"), logical connectives ("or" vs. "and"), and numerals ("one" vs. "two") (Horn, 1972). *Ad-hoc* implicatures are contextually weak descriptions that negate stronger interpretations (e.g. "I did my math homework" implies that I did *only* my math, and not also my history homework).[1] While scalar and ad-hoc implicatures are similar in nature and simple for adults, they are often challenging for children. We investigate factors influencing children's pragmatic inferences across these types of descriptions.

Children's processing of scalar implicatures is a focal case study for pragmatic development. Although adults spontaneously compute scalar implicatures along lexical scales like <SOME, ALL>, children's performance on these scales is variable even fairly late in development (Noveck, 2001). Paradigms that require children to make truth judgments for complex propositions may underestimate their pragmatic abilities, however, and 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 an 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.

Overall, research on children's abilities to compute scalar implicatures indicates that their fragile performance may have less to do with their 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.

Children's performance in implicature tasks increases when they have stronger access to lexical alternatives, supporting this hypothesis. 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 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, Goodman, & Frank, 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. Never-

[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.

theless, work to date has varied widely in the particular scales, tasks, and measures that were used, and the developmental samples are relatively small while spanning several years of age. These concerns make it difficult to interpolate across findings and draw strong inferences about contrasts between contextually-supported (ad-hoc) and lexicalized (scalar) implicatures. Our current study aims to fill this gap.

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 instructions children heard across trials (ad-hoc vs. scalar descriptions crossed with implicature vs. unambiguous control targets), 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 replaced the ad-hoc trials so that all of the descriptions used scalar quantifiers. 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, 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. Overall, our findings suggest that scalar implicatures are difficult for preschoolers even in supportive contexts, and that stronger recognition of lexical alternatives boosts performance.

Experiment 1

Methods

Participants A planned sample of 48 children was recruited from Bing Nursery School at Stanford University. Participants were recruited from two age groups: 24 4.0- to 4.5-year-olds (M = 4;2) and 24 4.5- to 5.0-year-olds (M = 4;7). Two additional children were excluded for stopping the study early, and one was excluded due to experimenter error.

Stimuli Children were shown printed images of three book covers, each depicting four familiar items (see Figure 1). An initial training trial featured a single unique item on each cover. For each of the 18 test trials, one book contained four items of a kind (e.g. four dogs), one book contained a different set of four items (e.g. four cats), and one book contained two pictures of a new set and two pictures repeated from one of the other covers (e.g. two cats and two birds). Each set of books featured a different set of familiar items.

Procedure Participants were tested individually in a quiet room at their preschool. The experimenter explained that they would be playing a game in which she would think about one of the three books on each page and give a clue about it. She emphasized that she would only give one clue for each set, so children were to make their best guess about which book she was describing based on that clue. A breakdown of the trial types and sample scripts is provided in Table 1.

Children began the task with a training trial in which each of the covers featured a single unique item. Following this initial trial, children saw 18 test trials with new sets of familiar items. Children were instructed to point to the book they thought the experimenter was describing. If children pointed to more than one book or their response was unclear, they were reminded that the experimenter was talking about just one book, and were asked to touch the one book they thought she meant.

For ad-hoc trials (eight total), the experimenter described the target using names of the images pictured. Ad-hoc control trials referred to unambiguous targets (e.g. "On the cover of my book, there are dogs/birds" in Figure 1). Ad-hoc implicature trials required an inference about the speaker's intended meaning: e.g. "On the cover of my book, there are cats" could potentially refer to either the book with only cats or the book with cats and birds, but the speaker's decision to describe only cats suggests that she is referring to the cover with all cats and no birds, or else she would have mentioned both types of animals.

For scalar trials (ten total), the experimenter described the target using quantifiers. Scalar control trials referred to unambiguous targets using *all* or *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 scalar implicature trials, the experimenter used a weak quantifier in reference to the item pictured across two book covers (e.g. "On the cover of my book, some of the pictures are cats"). Because the speaker used a weak quantifier, the implicature is that she must mean the cover with two cats and two birds, because if she had meant the cover with only cats, she would have use the stronger quantifier (*all*) instead.

Image sets were presented in a fixed order, counterbalanced for target location and book triad positions. The description condition (ad-hoc or scalar quantifier) and trial type (implicature or control) for each book set were randomized across participants. The conditions (ad-hoc or scalar) and trial types (implicature or control) were spaced as much as possible so that two trials of the same type never occurred twice in a row. Children enjoyed the task, responded quickly to the clues, and often made statements such as, "I'm good at this!" although they were not provided feedback about their selections. The test session took about ten minutes to complete.

Results

Children's performance on all trial types is shown in Figure 2 (responses were coded as correct on implicature trials if children chose the implicature-consistent target). Across all of the ad-hoc trial types, children were near ceiling in selecting

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

Condition	Trial type	# trials in Expt. 1	# trials in Expt. 2	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

the intended target. Using a novel task, our finding replicates previous work indicating that preschoolers can compute ad-hoc implicatures (Stiller et al., 2014), and suggests that children can make such inferences even in the presence of varied types of descriptions (control trials and scalar references).

Children's performance on scalar trials was markedly different and much lower. We ran a logistic mixed effect model, predicting correct responses as the interaction of age, condition (ad-hoc or scalar) and trial type (implicature or control), with random effects of participant and trial type. Performance was marginally lower for scalar trials than ad-hoc trials ($\beta = -8.02$, $p = .09$), and there was a significant interaction between condition and trial type, such that performance was significantly worse on scalar implicature trials ($\beta = 16.45$, $p = .02$). There was also 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.

Although children largely made correct choices on the *all* trials, their responses were more varied for *some* and *none* trials. Examining 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* ($D = .20$, $p < .0001$) trials, indicating that individuals tended not to respond at chance, but either consistently correctly or incorrectly on these trials. Additionally, children's success on *some* and *none* trials was highly correlated[2] ($r = .47$, $p < .001$) such that children who performed better on *some* trials also tended to perform better on *none* trials (see Figure 3). Performance on *none* and *all* trials ($r = .11$, $p = .45$) or *some* and *all* trials ($r = .01$, $p = .95$) was not correlated.

Discussion

Overall, we found that scalar implicatures were hard for children in our task. We wondered why this difficulty might be the case, given that we had tried to reduce as many task demands as possible in our task. Despite the presence of both visual alternatives (via the three selection choices) and lexical alternatives (conveyed across trials), children were at chance in their selections on scalar implicature trials.

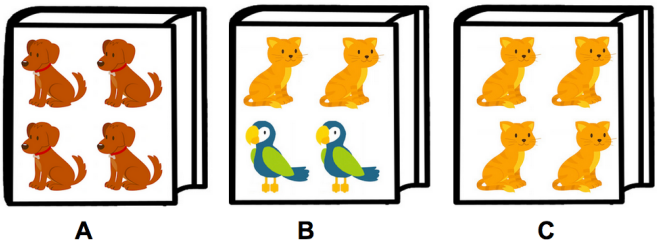

Figure 1: Example trial image set. Children saw three book covers with familiar images. The experimenter provided a clue about which of the three covers she was thinking of, using either an ad-hoc or scalar description of either an unambiguous or implicature target (see scripts in Table 1).

We also found an un-predicted developmental change in responses on *none* trials, corroborating recent work indicating that even older preschoolers show difficulty in the comprehension of negation in arbitrary contexts (Nordmeyer & Frank, 2014). We had expected *none* trials to serve as a simple unambiguous control, but found that this term was difficult for children (and that success was correlated with implicature performance). Perhaps children's implicature perfomance depends to some degree on familiarity with the both ends of the quantifier scale (*none—some—all*). They may need to recognize both extremes of the scale before consistently identifying the meaning of the intermediate, *some* term. Alternatively, another mediating factor (e.g., inhibitory control) might be responsible for the correlation we observed. We return to this issue in the General Discussion.

Finally, although our goal was to examine pragmatic development by comparing children's performance across a variety of inference trials, we wondered if including *both* ad-hoc and scalar quantifier descriptions led children to form expectations about the speaker that influenced their responses. We were concerned that their overwhelming success on ad-hoc

[2]This correlation was also replicated in a pilot version of this task, n=22, $r = 0.94$, $p < .0001$.

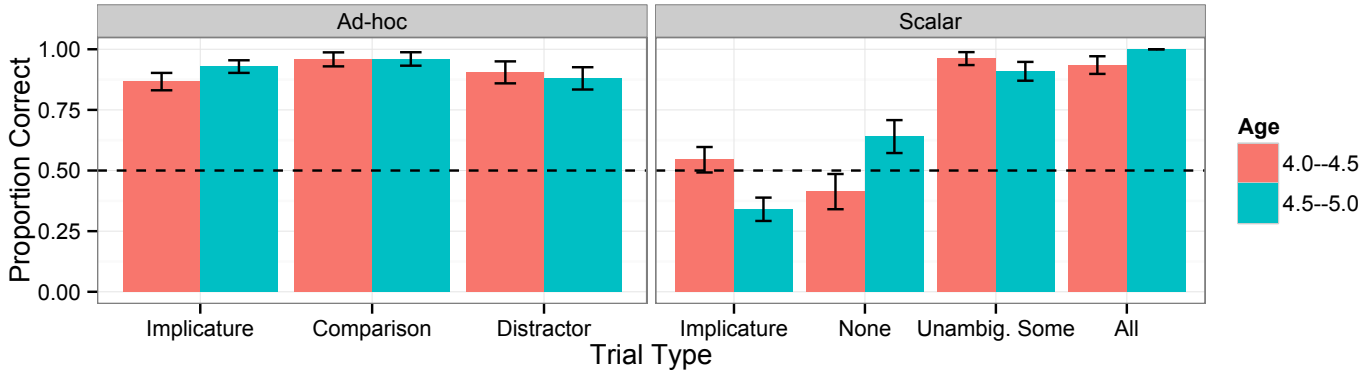

Figure 2: Proportion of correct target responses by age group for ad-hoc and scalar conditions in Experiment 1. Error bars show standard error of the mean.

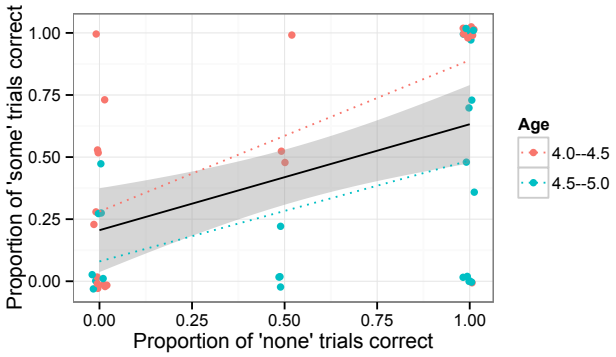

Figure 3: Scatterplot relating individuals' performance on *some* and *none* trials per age group in Experiment 1. The aggregate trend is plotted in black along with its 95% confidence interval. Trends for each age group are shown by the dotted lines. Points are jittered slightly to avoid overplotting.

implicature trials (e.g. "On the cover of my book, there are cats") might lead them to misinterpret the intention of *some* in scalar implicature trials from "... some of the pictures are cats" to "... there are some cats." If children are forming expectations about the speaker that override their sensitivity to the particular word choices in the referential expression, then their performance may be biased by the presence of the ad-hoc trials. To investigate this idea, we removed ad-hoc trials and ran a scalar-only version of the study.

Experiment 2

To investigate whether preschoolers would show increased sensitivity to individual quantifier use in the absence of competing ad-hoc descriptions, we ran a version of Experiment 1 using only scalar quantifiers. Additionally, in order to explore the developmental course of scalar implicature comprehension, we extended our sample to span the age range from 3–5 years, broken into half-year age groups.

Methods

Participants We recruited a new sample of participants from Bing Nursery School: 12 3.0–3.5 year-olds (M=3;4), 12 3.5–4.0 year-olds (M=3;8), 14 4.0–4.5 year-olds (M=4;3), and 12 4.5–5.0 year-olds (M=4;8). One additional child was excluded for stopping the task early.

Stimuli The same materials were used as in Experiment 1. The only changes made were to the scripts, such that ad-hoc trials were removed and all trials were converted into scalar quantifier references (Table 1). The 18 test trials contained six control *all* trials (e.g. "On the cover of my book, all of the pictures are cats"), six control *none* trials ("On the cover of my book, none of the pictures are cats"), and six scalar implicature *some* trials ("On the cover of my book, some of the pictures are cats"). Image sets were presented in a fixed order, counterbalanced for target location and triad order. Participants were randomly assigned to one of three scripts, with a pseudo-randomized trial order such every book set was referred to by each quantifier type (*all, none* or *some*), and the same trial type never occurred twice in a row.

Procedure The procedure was identical to Experiment 1.

Results and Discussion

Children's performance increased with age for all trial types (Figure 4). All age groups were strongest on the *all* trials, and the oldest children (4.5–5.0 year-olds) were the only age group above chance for *none* trials ($t = 3.09$, $p = .01$); this group was marginally above chance for *some* trials ($t = 1.85$, $p = .09$).

We ran a logistic mixed effect model, predicting correct responses as the interaction of age and trial type (*none, some* or *all*), with random effects of trial type by participant. Surprisingly, the only significant effect that emerged was age, such that performance increased across trials as children got older ($\beta = 20$, $p < .001$). We added trial order (first or second half

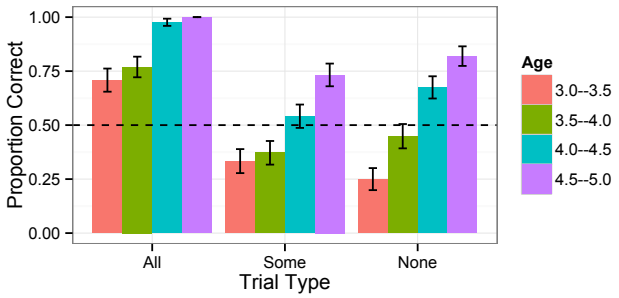

Figure 4: Proportion correct target responses per age group for each of the scalar trial types (*all, some* and *none*) in Experiment 2. Error bars show standard error of the mean.

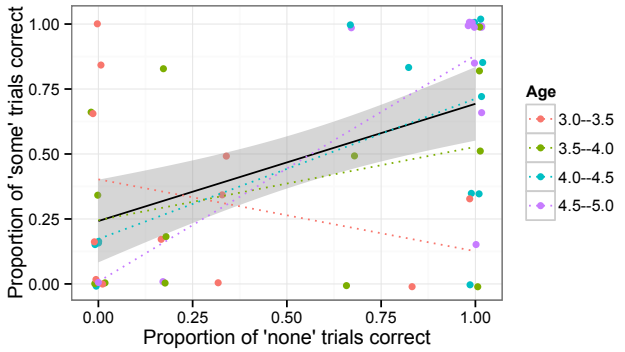

Figure 5: Scatterplot of individuals' performance on *some* trials and *none* trials in Experiment 2 (main effect in black, correlations per age groups illustrated by the dotted lines). Points are jittered slightly to avoid overplotting.

of the experiment) to the model but it did not interact with any variables, indicating that performance did not change over the course of the experiment. The lack of trial type effects in the main model was caused by the participant-level random effects structure and suggests that trial-type effects were not stable across participants (Barr, Levy, Scheepers, & Tily, 2013).

Consistent with the findings from the mixed effects model, we again found significant bimodal patterns of responses for both *some* ($D = .12$, $p < .0001$) and *none* ($D = .15$, $p < .0001$) trials. And again, these trial types were highly correlated with one another ($r = .52$, $p < .001$; Figure 5).

As an exploratory analysis, we ran another version of the mixed effects model removing the random effect of trial type. In addition to a main effect of age ($t = 1.88$, $p < .01$), this model revealed that performance on *some* trials was lower than *all* trials ($t = -7.69$, $p < 01$), and marginally reduced from *none* trials ($t = 3.03$, $p = .09$). We also found 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, children's success in selecting the speaker's intended target increased as children got older. Our results do not allow a strong inference about the cause of this developmental change, but several hints were present in the data. First, the bimodal and correlated patterns of responses for *none* and *some* suggests that children's knowledge of the full quantifier paradigm is not yet adult-like in their preschool years. One possible explanation is that they are learning that both *none* and *all* contrast with *some* in parallel. Second, there was a notable contrast between performance on *some* trials in Experiment 1 and Experiment 2, indicating that the presence of other (ad-hoc) trial types likely decreased children's implicature computation in our first experiment, and supporting the idea that pragmatic competition extends beyond the specific lexical scale being used (Degen & Tanenhaus, 2014).

General Discussion

We designed a simple task to test children's sensitivity to a variety of word choice cues in a single paradigm, allowing us 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. Our findings suggest that 4-year-olds are able to compute scalar implicatures, but their performance is fragile and reliant on contextual cues.

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 of the speaker's description. Second, the relatively high number of trials both helped strengthen our analytical power and also offered the possibility for children to identify lexical alternatives as the study progressed (although we did not find that their performance significantly changed over the course of either experiment, cf. Skordos & Papafragou, 2014). Third, we were able to not only compare performance across age groups, but also 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.

Our findings support 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 they are 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. Overall, these patterns of 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.

The correlated responses for *none* and *some* trials in both Experiments 1 and 2 present an interesting puzzle. *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 of (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. More research will be required to distinguish these possibilities.

One pattern in our data is 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).

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. Their computation of scalar implicatures increases in supportive contexts, but their inferences are fragile and depend on their knowledge of lexical items.

Acknowledgments

Special thanks to the Bing Nursery School, and Sara Altman and Carson Kautz for their help with data collection.

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 children's pragmatic inference. *Cognition*, *118*(1), 84–93.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Degen, J., & Tanenhaus, M. K. (2014). Processing scalar implicature: A constraint-based approach. *Cognitive Science*, 1–44.

Grice, H. (1975). Logic and conversation. *1975*, 41–58.

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*, 67–81.

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

Nordmeyer, A. E., & Frank, M. C. (2014). The role of context in young childrens 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*(1), 71–82.

Skordos, D., & Papafragou, A. (2014). Scalar Inferences in 5-year-olds: The Role of Alternatives. In *Proceedings of the 38th Annual Boston University Conference on Language Development. Somerville, MA: Cascadilla Press.*

Stiller, A. J., Goodman, N. D., & Frank, M. C. (2014). Ad-hoc implicature in preschool children. *Language, Learning & Development*, 1-15.