Cognition 153 (2016) 6–18

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Original Articles

Children's derivation of scalar implicatures: Alternatives and relevance

Dimitrios Skordos [a,*], Anna Papafragou [b]

[a] Department of Linguistics and Cognitive Science, University of Delaware, 125 E. Main Street, Newark, DE 19716, USA
[b] Department of Psychological and Brain Sciences, University of Delaware, 108 Wolf Hall, Newark, DE 19716, USA

ARTICLE INFO

Article history:
Received 30 June 2014
Revised 6 April 2016
Accepted 10 April 2016
Available online 21 April 2016

Keywords:
Scalar implicature
Relevant alternatives
Pragmatic inference
Pragmatic development

ABSTRACT

Utterances such as "Megan ate some of the cupcakes" are often interpreted as "Megan ate *some* but *not all* of the cupcakes". Such an interpretation is thought to arise from a pragmatic inference called *scalar implicature* (SI). Preschoolers typically fail to spontaneously generate SIs without the assistance of training or context that make the stronger alternative salient. However, the exact role of alternatives in generating SIs remains contested. Specifically, it is not clear whether children have difficulty with spontaneously generating possible informationally stronger scalemates, or with considering how alternatives might be relevant. We present three studies with English-speaking 5-year-olds and adults designed to address these questions. We show that (a) the accessibility of the stronger alternative is important for children's SI generation (Experiment 1); (b) the explicit presence of the stronger alternative leads children to generate SIs only when the stronger scalar term can easily be seen as relevant (Experiment 2); and (c) in contexts that establish relevant alternatives, the explicit presence of the stronger alternative is not necessary (Experiment 3). We conclude that children's considerations of lexical alternatives during SI-computation include an important role for conversational relevance. We also show that this more nuanced approach to the role of lexical alternatives in pragmatic inference unifies previously unconnected findings about children's early pragmatic development and bears on major accounts proposed to date for children's problems with SIs.

1. Introduction

1.1. Scalar implicatures

Implicatures are components of speaker meaning that constitute an aspect of what is *meant* in a speaker's utterance without being part of what is *said*. A *scalar implicature* (SI) is a pragmatic inference triggered by certain lexical items such as quantifiers. Often, the use of a proposition containing a quantifier such as *some* is taken to implicate that another proposition containing a logically stronger quantifier (*all*) would not hold. For example, the statement in (2a) below can be used to implicate (2b).

(2) a. Megan ate some of the cupcakes.
b. Megan did not eat all of the cupcakes.

The term *scalar* comes from the fact that linguistic terms like *some* and *all* form an ordered set of alternatives (a *scale*) based on informational strength[1] (<*all*, …, *most*, *some*, >; Horn, 1972). Informational strength is based on asymmetrical logical entailment where a proposition containing the informationally stronger term (*all*) logically entails a proposition containing the weaker one (*some*) but not vice versa.

On this account, the quantifier *some* has lower-bounded semantics ('at least some and possibly all'; Horn, 1972). The upper-bounded meaning ('some but not all') corresponds to the scalar implicature and is therefore a pragmatic enrichment of the semantic content of the quantifier. The conclusion that the upper-bounded meaning is a pragmatic, not a semantic, contribution is further supported by the fact that this meaning can be explicitly canceled without logical contradiction ("Megan ate some of the cupcakes. In fact, she ate all of them"). Other logical scales are based on logical connectives (<*or*, *and*>) or modals (<*might*, *must*>).

* Corresponding author.
E-mail addresses: dskordos@udel.edu (D. Skordos), apapafragou@psych.udel.edu (A. Papafragou).

http://dx.doi.org/10.1016/j.cognition.2016.04.006
0010-0277/

[1] There are newer re-interpretations of the original notion of Horn scales (e.g., Geurts, 2010) that view scales as a way to restrict scalar alternatives to what is relevant. However, the original and still widely used Horn (1972) notion of scales had very little (if anything) to do with relevance: scales and scalars were purely based on informativeness/quantity, with an allowance perhaps for the Quality maxim (see Matsumoto, 1995 for detailed discussion).

For instance, the statements in (3a) and (4a) below can be taken to implicate (3b) and (4b) respectively.

(3) a. Megan ate a cupcake or a cookie.
b. Megan did not eat both a cupcake and a cookie.

(4) a. Bert might be in his lab.
b. It is not the case that Bert must be in his lab.

Scalar implicatures can also be derived from non-logical scales, based on contextual information (Hirschberg, 1985). In some sense the terms "scales" and "scalar" are actually a misnomer: As Hirschberg has convincingly shown (1985) any partially ordered set can give rise to SIs. For instance, the response in (5b) implicates that the action of changing the oil was not completed.

(5) a. Did you change the oil?
b. I opened the hood.

The first account of how scalar implicatures are derived was described by Paul Grice. He suggested that communication is a co-operative effort largely governed by rational expectations about how a conversation should proceed. These expectations were formalized as a number of principles or maxims that are thought to guide the inferences which hearers usually entertain when interpreting utterances (Grice, 1975). When these expectations seem to be violated, the assumption that this was done on purpose creates a variety of effects (see also Horn, 1972). For instance, in (2a), the speaker has violated the Quantity maxim that asks speakers to make their contribution as informative as is required by the current conversational purposes: *some* is the less informative term within the scale <*some*, *all*>. Thus the choice of the weaker term *some* is reason to believe that the speaker cannot commit to an informationally stronger statement ("Megan ate all of the cupcakes."). Therefore, the stronger statement does not hold, thus (2b).

1.2. How children calculate SIs

The psycholinguistic literature has shown that adults are very adept at deriving scalar inferences (e.g., Bott, Bailey, & Grodner, 2012; Breheny, Ferguson, & Katsos, 2013; Breheny, Katsos, & Williams, 2006; Huang & Snedeker, 2009a). Young children, however, seem to face difficulties. For instance, Noveck (2001) showed that French speakers between the ages of 5 and 10 interpreted the French existential quantifier *certains* ("some") in statements such as "Some giraffes have long necks" as compatible with *tous* ("all"), while adults were equivocal between the logical and the pragmatic interpretations. Similarly, in another study, Greek-speaking 5-year-olds, unlike adults, accepted statements such as "Some of the horses jumped over the fence" as descriptions of story outcomes where *all* of the horses in the scene jumped over the fence (Papafragou & Musolino, 2003).

Subsequent studies have replicated and confirmed the finding that children typically display non-adult behavior when interpreting scalar statements (Feeney, Scrafton, Duckworth, & Handley, 2004; Foppolo, Guasti, & Chierchia, 2012; Guasti et al., 2005; Katsos & Bishop, 2011; cf. also Braine & Rumain, 1981; Smith, 1980). Importantly, children's difficulties emerge even in studies that used eye movement measures, as opposed to overt pragmatic judgments, to gain insight into comprehension (Huang & Snedeker, 2009b). Furthermore, a variety of factors seems to affect children's success with scalar implicatures. These include training in detecting pragmatic infelicity and/or a strong supporting context (Foppolo et al., 2012; Guasti et al., 2005; Papafragou & Musolino, 2003; Papafragou & Tantalou, 2004); the type of scale (logical vs. ad hoc; Barner, Brooks, & Bale, 2011; Stiller, Goodman, & Frank, 2015) and scalar item (number vs. quantifier; Papafragou, 2006; Papafragou & Musolino, 2003); and the type of response children have to provide (Katsos & Bishop, 2011; Pouscoulous, Noveck, Politzer, & Bastide, 2007; see Papafragou & Skordos, 2016, for a review).

Several strands of evidence suggest that part of children's problem with SIs lies in generating scalar alternatives when faced with a weak scalar term. In early studies that examined the interpretation of the disjunction operator *or* (Chierchia, Crain, Guasti, Gualmini, & Meroni, 2001; Gualmini, Crain, Meroni, Chierchia, & Guasti, 2001), adults were shown to be sensitive to the scalar implicature from the use of disjunction: when faced with statements like "Every boy chose a skateboard or a bike" to describe the outcome of a story, adults tended to interpret the statement as meaning 'Every boy chose either a skateboard or a bike'. However, 3–5-year-old children seemed oblivious to this pragmatic interpretation and treated *or* as being compatible with the stronger term *and*. In a follow-up task, however, children were presented with two statements and they overwhelmingly preferred stronger/more informative statement with *and* ("Every farmer cleaned a horse and a rabbit") over the weaker/less informative statement with *or* ("Every farmer cleaned a horse or a rabbit") when the story made the stronger statement true. Thus children could compare alternatives to a weak scalar term and assess their relative informativeness when these alternatives were explicitly presented to them but did not seem to independently access those scalar alternatives and use them to compute implicatures (see also Ozturk & Papafragou, 2015, for similar results with epistemic modals such as *may* and *have to*).

A study by Barner et al. (2011) offers further evidence for the role of the accessibility of unspoken lexical alternatives on children's SI calculation. Barner et al. tested 4-year-old children in a task that involved answering questions about a group of three animals. In critical trials, all three animals (a dog, a cat and a cow) were sleeping and children were asked whether "...some/only some of the animals are sleeping". Children responded affirmatively about 66% of the time even to the question with *only some*. This was taken to indicate that children have difficulty generating scalar alternatives even when this is predicted to be triggered by the grammar (*only* is a focus element requiring the generation and negation of appropriate alternatives). However when a different group of children were simply asked whether "the cat and the dog are sleeping", children accurately responded with an affirmative answer 93% of the time. More importantly, when asked whether "only the cat and the dog are sleeping", children correctly gave *No*-responses 86% of the time. Barner et al. (2011) interpreted these findings as strong evidence that children's problem with SIs lies mainly in realizing what terms can come together to form a scale: when scalemates are explicitly provided (e.g., when the experimenter listed the animals that were supposed to be sleeping), children's generation of SIs improved significantly.

Even though these studies suggest that the accessibility of scalar alternatives contributes to children's difficulties with SIs, the precise role and potency of lexical alternatives in the derivation of SIs at present remain open. One issue is that children's apparent insensitivity to SIs has been found even in contexts that should make stronger scalar alternatives highly accessible. For instance, in Noveck's (2001) judgment study, the critical true but infelicitous *some*-statements (e.g., "Some giraffes have long necks") were embedded within a larger battery of statements that also included other types of *some* statements and a variety of *all* statements (e.g., "All elephants have trunks"): even though this paradigm presumably made the stronger scalar alternatives accessible, children did not seem to benefit from the presence of the stronger term. In another study, when 5-year-olds were asked to evaluate an underinformative *some*-statement accompanying a story (e.g., "Some

smurfs went on a boat" in a story where all smurfs had gone on a boat), they were no better or worse at detecting underinformativeness compared to another task in which the *some*-statement was presented after children had to evaluate a true *all*-statement (e.g., "All of the dwarfs are eating a piece of candy"): in both cases, 5-year-olds gave largely logical responses when judging pragmatically infelicitous statements (Foppolo et al., 2012, Exp.3). These findings challenge the hypothesis that the difficulty of spontaneously generating stronger scalar terms is a major factor in children's computation of SIs.

A related issue is that, to the extent that the accessibility of stronger lexical alternatives facilitates SI computation in children, the mechanisms whereby this effect is achieved are not well understood. One possibility is that children have problems retrieving the stronger alternative when required. This possibility is more naturally aligned with accounts on which lexical scales containing quantifiers, modals and similar expressions provide a pre-defined set of alternatives that feed into the computation of at least certain classes of SIs (what have been termed 'generalized' SIs; see Chierchia, 2004; Chierchia, Fox, & Spector, 2009; Levinson, 2000). A particular example of this general class of accounts is the "restricted alternatives" hypothesis proposed by Tieu, Romoli, Zhou, and Crain (2015). In the words of Tieu et al.: "Take the case of scalar quantifiers. The child must learn that *some* and *all* lie on the same quantifier scale. A failure to compute the implicature could arise either because the child has yet to learn the co-scalar status of *some* and *all*, or because the child is unable to retrieve *all* from the lexicon during the experiment. [...]." (ibid., p.25). A related, albeit somewhat broader proposal is the "processing limitation" hypothesis (Chierchia et al., 2001; Gualmini et al., 2001). On this hypothesis, the process of computing the semantic content of an utterance, generating alternatives to the weak scalar term and rejecting them to strengthen the original proposition pragmatically might overwhelm younger children's processing abilities because of "the processing cost associated with maintaining in memory different representations of the target sentence" or "... involved in comparing different alternative representations of a sentence" Chierchia et al., 2001 p. 167). These general types of account (that we will collectively refer to as the "lexical retrieval" account for brevity) predict that, other things being equal, making the stronger lexical alternative available (e.g., by mentioning it explicitly prior to the weak scalar) should lead children to derive SIs from weak scalars.

Alternatively, children may not have a problem with considering possible lexical members of scales per se (or spontaneously activating them); the problem might lie rather in their failure to recognize that the scalar terms constitute *relevant* alternatives. For instance in discussing children's apparent insensitivity to SIs in judgment tasks, Papafragou and Musolino (2003) observed: "If preschoolers, unlike adults, cannot readily infer the pragmatic nature of the task, and are not given adequate motivation to go beyond the truth conditional content of the utterance, they may readily settle for a statement which is true but does not satisfy the adult expectations of relevance and informativeness" (p.269). They went on to propose that "if children are provided with a context where communicative [i.e., relevance] expectations are clear and where the stronger alternative to the weaker statement is made particularly salient, they will be more prone to noticing the implicature" (ibid, p.277; see also Foppolo et al., 2012, Exp.6; Bott & Noveck, 2004; Noveck & Sperber, 2007; cf. also Pouscoulous et al., 2007, where it is hypothesized that children's problems with SIs are due to difficulties in the optimization process between cognitive gains and processing costs as defined within a relevance-theoretic framework.). Such an account would predict that simply activating the stronger lexical item will not necessarily be enough for children to derive a SI, unless children appreciate that this alternative to the statement offered (out of many different possibilities) is relevant to the goal of the conversation.

Existing developmental studies cannot adjudicate between these two possibilities about the role of lexical alternatives because they have not independently manipulated the contribution of the accessibility vs. relevance of lexical alternatives in children's SI generation. Our goal is to do so in the studies that follow. Clarifying the role of alternatives in the computation of SIs has broader implications for further theories of children's pragmatic difficulties with SIs that do not attribute a particular role to the accessibility of scalar alternatives (Katsos & Bishop, 2011; Noveck, 2001). We discuss these theories more fully towards the end of the paper.

1.3. The present studies

In the present studies we explored the role of lexical alternatives in children's computation of scalar inferences. Our main goal was to throw light on both the potency of lexical alternatives on children's derivation of SIs and the theoretical machinery whereby such alternatives exert their effects.

We focused on the quantificational scale <*some*, *all*>. In Experiment 1, we tested whether the presence of the stronger lexical member of the quantifier scale (*all*) in the course of the experiment can encourage children to generate a SI from the use of a weak alternative (*some*). If considering stronger scalar candidates is a limiting factor in children's computation of SIs, then children's pragmatic performance should improve when the stronger scalar term (*all*) is provided for them.

In Experiment 2, we explored the nature of the mechanism that uses lexical alternatives to generate SIs. If lexical alternatives lead to SI generation simply by virtue of providing children with the stronger lexical scale member that they typically fail to consider or activate, then children's computation of SIs should improve when *all* is provided for them regardless of whether *all* is shown to be a relevant scalar alternative to the weak scalar term used or not. If, however, the relevance of lexical alternatives plays a role in SI generation, then the availability of the stronger alternative *all* should have an effect on children's computation of SIs only if the lexical scale member can be shown to be a relevant alternative.

Finally, in Experiment 3, we asked whether the generation of SIs from *some*-statements in children can be achieved even in the lexical absence of the stronger scalar term *all* if another quantifier (e.g., *none*) can be used by children as a cue to access the scale <*some*, *all*>. This prediction is unexpected on lexical retrieval accounts, according to which supplying the stronger scalar alternative (*all*) should be a privileged way of helping children recover otherwise inaccessible scalar structure.

2. Experiment 1

In Experiment 1 we introduced the basic paradigm that appears throughout the present studies. We used an Acceptability Judgment Task (AJT) similar to those in prior work (Chierchia et al., 2001; Foppolo et al., 2012; Guasti et al., 2005; Papafragou & Musolino, 2003). In our task, scalar terms (*some* and *all*) were embedded within statements that needed to be evaluated based on visual evidence in the scene. Critical trials designed to assess children's generation of scalar inferences consisted of true and infelicitous *some* statements that needed to be rejected if one derived a scalar implicature. We manipulated the accessibility of the stronger alternative through the order of the *some* and *all* statements. Of interest was whether providing the stronger lexical member of the scale would affect children's ability to generate a SI from the critical *some*-statements.

Our design had two noteworthy differences from prior studies. First, we included semantic controls to test for children's understanding of the semantics of the quantifiers. Even though there is evidence that children show some understanding of *all* and *some* from around the age of two, children's use of *some* is not really consistent until the age of 5 and even then it is not completely error free (Barner, Chow, & Yang, 2009). To be able to examine the pragmatic competence of children who have already mastered the semantics of the quantifiers, in some of our analyses we used the semantic trials as controls to exclude participants whose semantic performance was low. Second, we restricted the universe of discourse for quantified phrases so as to make the evaluation of quantified statements more transparent: quantifiers always ranged over a unique set of 4 novel creatures ("blickets") such that the evaluation of the statements was based only on the visual context provided by each trial.

2.1. Method

2.1.1. Participants

We tested 90 typically developing 5-year-old children (4;10 – 5;11, M = 5;3) and 36 adult controls, all monolingual speakers of English. The children were recruited from daycare centers in Newark, DE. The adults were college students recruited from the University of Delaware, and received course credit for their participation. An additional group of 7 children were tested but excluded from the analysis for failure to follow instructions (n = 3) or for misidentifying objects in the displays as made evident by their responses (n = 4).

2.1.2. Materials and procedure

Children sat in front of a laptop PC computer and were shown the slides depicting the experimental stimuli. A first experimenter introduced the task to the children by introducing a hand-held puppet, Max the silly gorilla, "who says silly things sometimes", and explaining that they would see some pictures on the computer together. Participants were told that the puppet would describe the pictures and that they would have to say whether the puppet "said it well or not". They would also have to justify their answer in case they rejected the puppet's statement. A second experimenter animated the puppet and provided the appropriate statements, while the first experimenter wrote children's answers down on an answer sheet. Adults were tested in a very similar way, with the only differences being that (a) they had to write down their own responses in answer sheets, with the options *Yes/No* and space to justify *No*-answers and (b) they were tested in groups without the presence of a puppet (they were shown a cartoon character, Max the silly gorilla, that supposedly provided the statements that the experimenter read out).

Participants first went through 4 pre-test trials. These consisted of slides depicting cartoon animals or objects (e.g., a cow, an ice cream cone). Two of the pre-test trials were erroneously described by the puppet and two of them were correctly described, so that participants would have evidence that the puppet was capable of providing both 'silly' and accurate statements. For pre-test trials, participants were also provided with feedback when they failed to reject a false statement. For example, if participants agreed with the puppet when it described the cow as an "elephant", the experimenter would explain that the puppet "didn't say it well', and that in fact the picture depicted a cow.

After the pre-test trials concluded, participants were introduced to a cartoon character, Ben the Wizard, presented on a new slide. Ben was shown to use his magic wand to create the 4 "blickets", novel animate creatures that would appear on all test slides. Participants were informed that these were "the only blickets in the whole world". The experimenter next introduced the main phase of the experiment: "Now we are going to play a game with Max the silly gorilla and the blickets. We are going to look at some pictures about the blickets on the computer, and Max is going to say something about the pictures again. Once more, I am going to ask you to tell me whether he said it well, or not, OK?"

For the main trials, blickets were paired with everyday items (e.g., crayons, flashlights, paintbrushes, etc.) to create 16 basic scenes. Each basic scene had two versions: in *full set* scenes, 4 out of 4 blickets would have the item, and in *partial set* scenes, 3 out of 4 blickets would have the item. For each basic scene, there was a corresponding statement with two possible quantifiers (*Some/All of the blickets have an X*) that could be offered by Max. For each of the 16 basic scenes, we crossed scene type (full vs. partial set) with statement type (*some* vs. *all*) to create 4 types of trials: *True-All*, *False All*, *True-Some* and *True-and-Infelicitous-Some* trials. The total of 64 trials was split across 4 different stimulus lists such that, for each basic scene, a different scene type was paired with a different quantifier in each list. Each stimulus list contained a total of 16 test trials (with 4 trials of each type). Each participant saw one stimulus list. Examples of trial types within a list are given in Fig. 1. Notice that *True-All*, *False All*, and *True-Some* trials tested participants' semantic judgments about *some* and *all*. *True-and-Infelicitous-Some* trials tested participants' pragmatic judgments (i.e., their ability to generate SIs): even though it was logically true that some of the blickets had the item mentioned, the statement was infelicitous because in fact *all* of them did.

The internal order of trials within each list was manipulated across 3 between-subjects conditions: In the *Mixed* condition, *some*- and *all*- trials were intermixed in a pseudorandom order such that trial type (*True-All*, *False All*, *True-Some* and *True-and-Infelicitous-Some*) alternated at least every three trials. Thus the stronger lexical scale member *all* was highly accessible during the evaluation of *some* statements (including the critical underinformative ones). In the *Some-First* condition, *some*- and *all*- trials were presented in blocks, with the *some*-block always first (trial order within blocks was pseudorandom: trial types such as *True-All* vs. *False-All* for the *all* block and *True-Some* vs. *True-and-Infelicitous-Some* for the *some* block alternated at least every 3 trials). In this condition, the stronger lexical scale member (*all*) was not made available to children prior to evaluating *some* (including *True-and-Infelicitous-Some*) statements. Finally, in the *Infelicitous Some-First* condition, the *some*-block of the previous condition was further split into two blocks, with *True-and-Infelicitous-Some*-trials always presented first as a block and *True-Some*-trials always last, as a further way of reducing contrast within the class of *some* trials. The *all*-block remained unchanged.

2.2. Predictions

If considering stronger scalar candidates is a factor contributing to children's difficulties with scalar inference, children's rejection of *True-and-Infelicitous Some*-statements should improve when the strong lexical scale member (*all*) is made available to children: that is, children's pragmatic performance should be better in the *Mixed* than in the *Some-First* or *Infelicitous-Some-First* condition (with performance being the weakest perhaps in the *Infelicitous-Some-First* condition where there is no contrast to other scalar statements whatsoever). Alternatively, if considering stronger scalar candidates does not affect scalar inference, then children's SI generation (i.e., rejection of *True-and-Infelicitous-Some*-trials) should be comparable across the three conditions.

No difference in children's performance between conditions is predicted for the semantic trials (*True-All*, *False-All*, *True-Some*). Finally, no difference in adult performance is expected between conditions for either the semantic or the pragmatic trials.

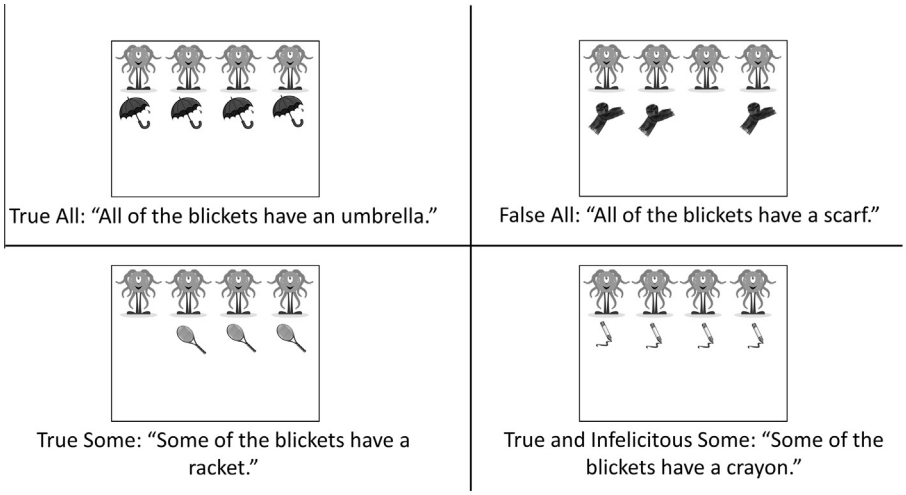

Fig. 1. Types of trials for Experiment 1.

Table 1
Participant performance in Experiment 1.

Trial type	Classification	Adults Condition			Children Condition		
		Mixed	Some-First	Inf-Some-First	Mixed	Some-First	Inf-Some-First
True-All	Passers	12	12	12	30	29	29
	Failers	0	0	0	0	1	1
False-All	Passers	12	12	12	25	24	24
	Failers	0	0	0	5	6	6
True-Some	Passers	12	12	12	26	26	23
	Failers	0	0	0	4	4	7
True-and-Inf-Some	Passers	12	12	10	23	14	7
	Failers	0	0	2	7	16	23

Note: The numbers represent Passers vs. Failers in corresponding trials.

2.3. Coding

Yes answers were coded as correct in the case of true statements. *No* answers were coded as correct in the case of false or true-and-infelicitous statements. A mean of correct answers from 0 to 1 was calculated for each participant for each of the 4 trial types (*True-All*, *False-All*, *True-Some*, *True-and-Infelicitous-Some*). Because 76 of the 90 children (or 84%) had scores of either 0 or 1 in the critical *True-and-Infelicitous-Some* trials, we categorized participants according to their performance on each trial type as either Passers (if they had a score of 0.75 or greater), or Failers (if they had achieved a score of 0.50 or less), and conducted non-parametric statistics on the data.

2.4. Results

Adult performance was at ceiling for all conditions and trial types (see Table 1). Fisher's Exact test analyses on 2×3 contingency tables for each trial type revealed no significant difference in the numbers of Passers vs. Failers across conditions (*True-All*-trials, $p = 1$; *False-All*-trials, $p = 1$; *True-Some*-trials, $p = 1$; *True-and-Infelicitous-Some*-trials, $p = 0.31$).

For children (see Table 1), Fisher's Exact Tests on 2×3 contingency tables did not reveal significant differences in the numbers of Passers vs. Failers across the 3 conditions for either the *True-All*-trials ($p = 1$), *False-All*-trials ($p = 1$), or *True-Some*-trials ($p = 0.52$). Turning to the *True-and-Infelicitous-Some*-trials, a Fisher's Exact test on a 2×3 contingency table revealed a significant difference ($p < 0.0001$) between the numbers of Passers and Failers across the 3 conditions. This effect was further explored by running Fisher's Exact Test on 2×2 contingency tables comparing each condition to the others. It was found that the *Mixed* condition had significantly more Passers than either the *Infelicitous Some-First* condition ($p < 0.0001$) or the *Some-First* condition ($p = 0.03$). There was no difference in the number of Passers between the *Some-First* and the *Infelicitous Some-First* condition ($p = 0.103$).[2]

When asked to justify their rejections of *True-and-Infelicitous-Some* statements, children overwhelmingly referenced either the stronger scalar term (e.g., "*All* of them/the blickets have an X"; 38 out of 44 Passers, or 86%), or mentioned the number of blickets and items available (e.g., "There is 4 blickets and 4 crayons"; 5 out of 44 Passers, or 11%). This shows that children rejected the *True-and-Infelicitous Some*-statements for the correct reason, namely because they generated the appropriate SI. The same holds true for adult participants (referenced the stronger scalar term: 33 out of 34 Passers, or 97%; mentioned the number of blickets and items available, 1 Passer, or 3%).

As is obvious from Table 1, some of the children performed poorly in the *False-All* and *True-Some*-trials. This raises doubts as

[2] A binary logistic regression run with Condition (Mixed, Some-First, Infelicitous-Some-First) as a predictor and children's performance in the *True-and-Infelicitous Some* trials (Passer, Failer) as a binary dependent variable, returned a main effect of Condition: *Wald's* $\chi^2(2) = 15.267$, $p < .0001$. This main effect was further explored by pairwise comparisons (Bonferonni correction) that revealed that the numbers of passers and failers in the *Mixed* condition were significantly different from those in the *Some-First* condition ($p = .036$) and those in the *Infelicitous-Some-First* condition ($p < .0001$). There was no difference in the numbers of passers and failers between the *Some-First* and the *Infelicitous-Some-First* condition ($p = .152$).

to whether these children have fully acquired the semantics of the quantifiers. If this is the case, it is not clear that one can derive conclusions about these children's pragmatic competence with quantifiers. To address this concern, we conducted a second analysis to exclude children who had under 0.75 correct in any of our control semantic trial types (either of the *True-All*, *False-All*, or *True-Some*-statements). This resulted in n = 8 children being excluded in the *Mixed* condition, n = 10 in the *Some-First* and n = 13 in the *Infelicitous-Some-First* condition.

This new analysis examined only performance on the *True-and-Infelicitous-Some*-trials in children who can be safely assumed to have the correct semantics for *some* and *all* (see Table 2). A Fisher's Exact test on the 2 × 3 contingency table in Table 2 revealed a significant difference between the numbers of Passers vs. Failers for the 3 different conditions ($p < 0.001$), confirming the results of the first analysis. This effect was further explored by running Fisher's Exact Test on 2 × 2 contingency tables comparing each condition to the others. Comparing the *Mixed* and the *Some-First* condition again revealed a significant difference ($p = 0.009$), with the *Mixed* condition having significantly more Passers than the *Some-First* condition. Comparing the *Mixed* and the *Infelicitous-Some-First* condition there was again a significant difference ($p = 0.0003$), with the *Mixed* condition having significantly more Passers than the *Infelicitous-Some-First* condition. Finally, comparing the *Some-First* condition to the *Infelicitous-Some-First* condition once again revealed no significant difference ($p = 0.33$).

Finally, the performance of the last group of children was compared with that of adults on the *True-and-Infelicitous-Some*-trials with Fisher's Exact Test on 2 × 2 contingency tables. There was no difference between age groups in the *Mixed* condition ($p = 1$), a significant difference in the *Some-First* condition, with the adult group having significantly more Passers than the child group ($p = 0.014$), and a trend towards a significant difference in the numbers of Passers vs. Failers for the *Infelicitous Some-First* condition ($p = 0.053$), with adults having more Passers than the child group.

2.5. Discussion

Experiment 1 was conducted to test the hypothesis that the presence of the stronger lexical scale member *all* would facilitate children's generation of SIs (Barner et al., 2011; Chierchia et al., 2001; Gualmini et al., 2001; Papafragou & Skordos, 2016). This hypothesis was supported by our data. In the *Mixed* condition, where *some*- and *all*-trials were intermixed so that the strong scalar term was made available to children by the time they had to judge the underinformative *True-and-Infelicitous-Some* statements, children were very successful at generating the appropriate scalar inference by rejecting infelicitous statements with *some*. Children's performance fell significantly when the stronger scalar term *all* was not provided for them (in the *Some-First* and *Infelicitous-Some-First* conditions). There was no difference in terms of detecting infelicity between the *Some-First* and *Infelicitous-Some-First* condition.

These results hold even when we look only at children that seem to have a solid grasp of the semantics of the quantifiers ('*some/all* knowers'). In the *Mixed* condition, these children's behavior is adult-like, unlike the conditions where the stronger scalar term is not present (*Some-First* and *Infelicitous-Some-First* condition). Overall, our results are consistent with prior evidence that 5-year-old children have difficulties with SIs, especially in judgment tasks (e.g., Guasti et al., 2005; Papafragou & Musolino, 2003), but their performance depends on the nature of the task (ibid.).

Even though the results of Experiment 1 provide evidence for the conclusion that the accessibility of the stronger scalar term plays a role in children's SI generation, the precise nature of children's difficulty with SIs and the way lexical alternatives helped them remain open. One possibility is that children could not spontaneously generate the stronger lexical scale member in order to access the scale. On this hypothesis, children's failures with SIs in the *Some-First* and *Infelicitous-Some-First* conditions in our study was a straightforward consequence of the fact that children could not access the scale without activation (initiated externally) of the lexical scale member. Alternatively, it may be that children did not always realize that a quantifier scale (<*some*, *all*>) could or should be accessed in the first place, i.e., no lexical terms could be seen as relevant scalar alternatives in the *Some-First* and the *Infelicitous-Some-First* condition. The results of Experiment 1 cannot adjudicate between these two possibilities. We therefore devised Experiment 2 to do so directly.

Table 2
Some/all-knowers' performance on *True-and-Infelicitous-Some* trials of Experiment 1.

Trial type	Classification	Children Condition		
		Mixed	Some-First	Inf-Some-First
True-and-Inf-Some	Passers	21	12	7
	Failers	1	8	10

3. Experiment 2

In Experiment 2 we further explored the role of scalar alternatives in the computation of SIs. Inspired by the *Mixed* condition of Experiment 1, we modified the Acceptability Judgment Task of our earlier study such that the block of *all* trials was always presented to children before the block of *some* trials. This ensured that the stronger lexical scale member (*all*) had been provided to the children (and therefore activated) when weaker (*some*) statements were encountered.

Within this set-up, we manipulated the degree to which the stronger lexical item could be easily recognized as a relevant alternative by children. This was accomplished by introducing subtle cues about the conversational goal (the evaluation criterion for the statements) in the first (*all*) block: In the Quantity condition, these cues pointed to the quantity of blickets that possessed an object as the relevant dimension for evaluating the statements. (e.g., Do *all* or *only some* of the blickets have a crayon?). In the Object condition, these cues pointed to the type of object possessed by the blickets as the relevant dimension for evaluating the statements (e.g., Do the blickets have *a crayon* or *another object*?).

This paradigm allowed us to test two different hypotheses about how lexical alternatives help children compute SIs that were left open in Experiment 1. Recall that, according to the lexical retrieval account, children have problems considering stronger lexical scale members, and thus, other things being equal, simply activating the necessary lexical scale members in an environment where SI generation is possible should lead children to derive SIs. According to an alternative account, however, simply providing stronger scalar terms should not necessarily lead children to derive SIs, since activation of the lexical scale members is not sufficient for children to see them as relevant competitors. Here the lexical retrieval account predicts that the degree of difficulty associated with recognizing a scalar term as a relevant alternative should not affect the SI computation when children encounter *some* statements in the second block of the study since activation of the stronger lexical scale member is guaranteed already from the first block in both conditions. By contrast, the account that posits an important role for conversational relevance predicts that 5-year-olds should benefit from the activation of the stronger lexical scale member in the present experiment only when this term can be easily seen as a relevant alternative to the weak scalar used (*some*). On this account, the Quantity condition should have an advantage

over the Object condition since, in order for children to succeed in rejecting true but underinformative *some* statements in the second block, they need to consider alternatives describing the quantity of blickets (*some* vs. *all*) that possess an object.

3.1. Method

3.1.1. Participants

We tested a group of 50 typically developing 5-year-old children (4;9 – 5;8, M = 5;0) and 24 adult controls, all monolingual speakers of English. None of these participants had taken part in Experiment 1. The children were recruited from daycare centers in Newark, DE. The adults were college students recruited from the University of Delaware and received course credit for their participation. An additional group of 4 children were tested but excluded from the analysis for failure to follow instructions (n = 3) or for misidentifying objects in the displays as made evident by their responses (n = 1).

3.1.2. Materials and procedure

The materials and procedure were very similar to those in Experiment 1 with the following major difference: the *all*-trials were always presented in a first block, and the *some*-trials in a second block so that lexical contrast between the stronger (*all*) and weaker (*some*) scalar terms could be established. Within each block, trials were presented in a pseudorandom order such that trial type (*True-All* vs. *False-All* for the *all* block and *True-Some* vs. *True-and-Infelicitous-Some* for the *some* block) would alternate at least every three trials.

There were two between-subjects conditions that differed only in the scenes accompanying the *False-All* trials within the *all* block: In the *Quantity* condition, the *False-All* statements did not match the quantity of the blickets in the visual scene, whereas in the *Object* condition, the same statements did not match the kind of object possessed by the blickets. For instance, in the *Quantity* condition, a statement such as "All of the blickets have a scarf" would be false because 3 out of 4 blickets would have a scarf but, in the *Object* condition, the same statement would be false because all 4 blickets would have a *shovel* (see Fig. 2; notice that the scenes for the *False-All* trials in the *Object* condition had to be modified slightly compared to Experiment 1 and the Quantity condition of the present experiment).

The rationale for this manipulation was the following: *False-All* trials uncover the dimension that the puppet was likely to err in (blicket quantity vs. object identity), and hence the basis upon which the participants were called to evaluate each statement. In the *Quantity* condition, this evaluation criterion (identifying the quantity of blickets having X) remained stable throughout the experiment: it was established in the first (*all*) block through the *False-All* trials and could later be brought to bear on judgments of the *True-and-Infelicitous-Some* statements. In the *Object* condition, however, the evaluation criterion changed between the first and the second block. The first (*all*) block, especially the *False-All* trials, should arguably lead participants to identify object identity as the evaluation criterion (i.e., whether the blickets possessed the stated object kind or not). In the second (*some*) block, however, if participants were to detect the infelicity of the *True-and-Infelicitous-Some* trials, they would have to use a different evaluation criterion (namely, whether the quantity of blickets in possession of a certain object was as stated in the sentence or not). We reasoned that, as long as the evaluation criterion remains stable between blocks, the corresponding *some*- and *all*-statements will be clearly seen as contrasting across the same dimension (quantity of the blickets possessing an item) and it will be easier for children to view the scalar terms *some* and *all* as relevant scalar alternatives. Conversely, if the evaluation criterion changes between blocks, the statements containing the scalar terms will not necessarily be seen as contrastive, since they are predicated across different dimensions (*all*-statements are predicated on object identity, while *some*-statements are predicated on blicket quantity). This should make it more difficult for children to view the scalar terms as relevant alternatives.

3.2. Predictions

If the lexical retrieval account is correct, then children should successfully reject *True-and-Infelicitous-Some* statements in both conditions, since *all* is lexically activated in both conditions by virtue of being present throughout the first block. However, if children have difficulties in viewing scalar terms as relevant alternatives, then children should be more successful in the *Quantity* than the *Object* condition, since the stronger lexical scale member *all* should only be seen as a relevant alternative in the *Quantity* condition where it is contrasted to the weak counterpart *some* across the same dimension (which should subsequently let children access the scale and compute the SI).

To see why this is so, consider the true but infelicitous statement "Some of the blickets have a crayon" (Fig. 2), uttered when all of the blickets have a crayon. If children believe that the goal is to evaluate whether the puppet got the quantity of blickets right (*Quantity* condition) and already have access to the stronger *all* term, they should easily reject the *some*-statement (since all of the blickets have a crayon). But if children believe that the goal is to evaluate whether the puppet got the object owned by the blickets right (*Object* condition), even if they have access to the stronger *all* term, they might not reject the statement (since some of the blickets indeed have a crayon).

Adult performance is not expected to differ between the two conditions as adult communicators should in principle be able to come up with the relevant alternatives and access the scale in order to derive the corresponding SI without help.

3.3. Coding

The coding scheme was identical to the one used for Experiment 1.

3.4. Results

Adult performance was very high for all conditions and trial types. Table 3 summarizes adult performance. Fisher's Exact test analyses on 2×2 contingency tables for each trial type revealed no significant difference in the numbers of Passers vs. Failers across conditions (*True-All*-trials, $p = 1$; *False-All*-trials, $p = 1$; *True-Some*-trials, $p = 1$; *True-and-Infelicitous-Some*-trials, $p = 1$). Adults were overwhelmingly pragmatic in the *True-and-Infelicitous-Some* trials.

Children performed well with the 3 semantic trial types (see Table 3). Fisher's Exact Tests on 2×2 contingency tables did not reveal significant differences in the numbers of Passers vs. Failers across the two conditions for either the *True-All*-trials ($p = 1$) or *False-All*-trials ($p = 0.357$), and only a difference approaching significance in the *True-Some-trials* ($p = 0.0504$). Turning to the critical *True-and-Infelicitous-Some*-trials, children appeared to be more pragmatic in the *Quantity* condition and more logical (non-pragmatic) in the *Object* condition: a Fisher's Exact test on a 2×2 contingency table revealed a significant difference ($p = 0.0465$) between the two conditions.[3]

[3] A binary logistic regression run with Condition (Quantity, Object) as a predictor and children's performance in the *True-and-Infelicitous Some* trials (Passer, Failer) as a binary dependent variable, returned a main effect of Condition: *Wald's* χ^2 (1) = 4.942, $p = .026$.

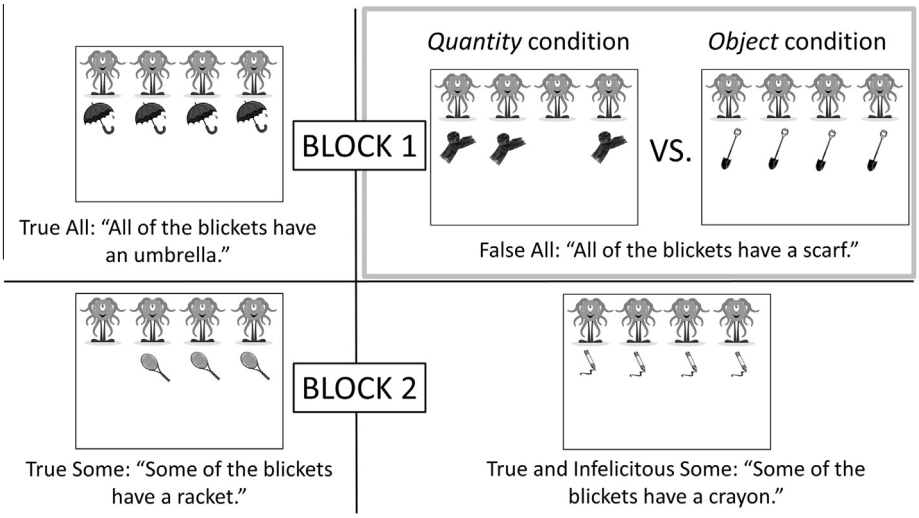

Fig. 2. Types of trials for Experiment 2. *False-All* trials differed between the *Quantity* and the *Object* condition.

Table 3
Participant performance in Experiment 2.

Trial type	Classification	Adults Condition		Children Condition	
		Quantity	Object	Quantity	Object
True-All	Passers	12	12	26	24
	Failers	0	0	0	0
False-All	Passers	12	12	22	23
	Failers	0	0	4	1
True-Some	Passers	12	12	19	23
	Failers	0	0	7	1
True-and-Inf-Some	Passers	11	12	17	8
	Failers	1	0	9	16

As in Experiment 1, when asked to justify their rejections of *True-and-Infelicitous-Some* statements, children referenced either the stronger scalar term ("*All* of them have an X", 22 out of 25 Passers, 88%), or the number of blickets that possessed an item ("Because there is 4 blickets and 4 X's" 3 out of 25 Passers, or 12%). Adults did similarly (reference to stronger scalar term 22 out of 23 Passers, 96%; number of blickets 1 Passer, or 4%).

After this initial analysis and for the same reasons as in Experiment 1, we conducted a second analysis excluding children who were Failers in any of our control semantic trial types (either the *True-All*, *False-All*, or *True-Some* statements). This resulted in n = 9 children being excluded in the *Quantity* condition, and n = 2 children excluded in the *Object* condition. All of these children can be assumed to have the correct semantics for *some* and *all*. A Fisher's Exact test on the 2×2 contingency table in Table 4 revealed a significant difference between the numbers of Passers vs. Failers for the two different conditions ($p < 0.0001$), with the *Quantity* condition having significantly more Passers than the *Object* condition, confirming the results of the first analysis.

We then compared the performance of the children who are *some/all*-knowers with that of adults with a Fisher's Exact Test on 2×2 contingency tables. No difference was found between age groups in the *Quantity* condition ($p = 0.414$), but we did find a significant difference in the *Object* condition, with the adult group having significantly more Passers than the child group ($p < 0.0001$).

Finally, to complete our analysis, we ran a Fisher's Exact Test on 2×2 contingency tables comparing *some/all*-knowers' performance in the *True-and-Infelicitous-Some* statements across the *Quantity* condition and each of the three conditions of Experiment 1. No difference was found between the *Quantity* and the *Mixed* condition ($p = 1$). However, significant differences were found in the numbers of Passers vs. Failers between the *Quantity* and the other two conditions (*Some-First*, $p = 0.004$; *Infelicitous-Some-First*, $p = 0.0003$) with the *Quantity* condition having significantly more Passers than either of the other two conditions. A similar comparison of the *Object* condition to the three conditions from Experiment 1 revealed a significant difference between the *Object* and the *Mixed* condition from Experiment 1 ($p < 0.0001$), with the *Mixed* condition having significantly more Passers than the *Object* condition. We found no difference between the *Object* condition and either the *Some-First* ($p = 0.06$) or the *Infelicitous-Some-First* condition ($p = 0.497$).

Table 4
Some/all-knowers' performance in *True-and-Infelicitous-Some* trials of Experiment 2.

Trial type	Classification	Children Condition	
		Quantity	Object
True-and-Inf-Some	Passers	17	6
	Failers	0	16

3.5. Discussion

Experiment 1 showed that at least part of the problem children face in SI generation lies in failing to spontaneously generate the stronger scalar term when a weak scalar term is used. Experiment 2 further explored this idea and asked whether the accessibility of the stronger lexical scale member could bear the explanatory burden of children's failure with SIs alone, or whether stronger scalar terms actually need to be seen as relevant alternatives in order to allow children to access a scale and derive a SI. Our results clearly support the second account: even in contexts where the stronger scalar term (*all*) was explicitly mentioned, children did not benefit from its presence unless the scalar term was seen as a relevant stronger alternative. Thus accessibility of the stronger scalar term is a necessary but not sufficient condition for the generation of SIs in children. These results cohere with and further clarify the findings from Experiment 1: specifically, they suggest that children's success with SIs in the *Mixed* condition of the earlier study was not simply due to the lexical availability of the stronger scalar term but was due to the fact that the stronger quantifier was seen as a relevant alternative.

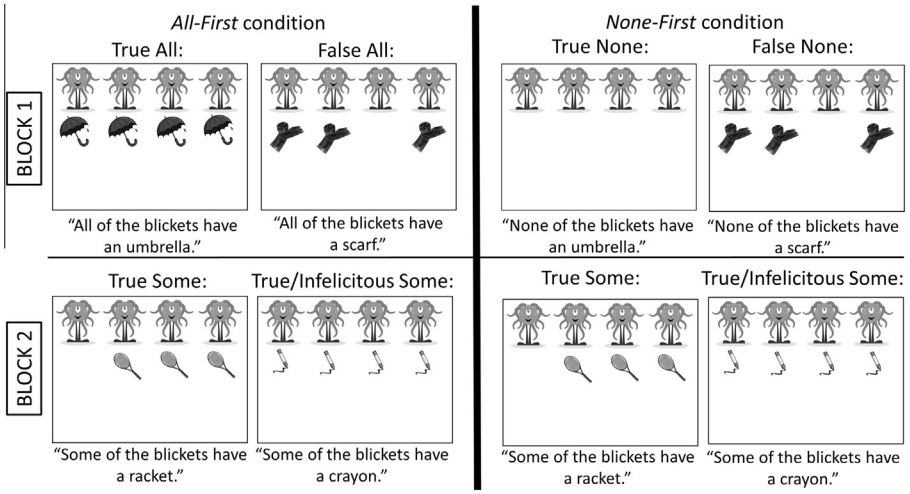

Fig. 3. Types of trials for Experiment 3.

4. Experiment 3

Experiment 2 showed that children benefit from the accessibility of the stronger scalar term only when the scalar term can be seen as a relevant alternative to another being used. Experiment 3 tested a further set of predictions from competing theoretical accounts of SIs. If children's difficulties with SI-computation relate to difficulty in generating stronger scalar terms, providing children explicitly with the stronger scalar term should provide a unique pathway into SI-computation. But if children need to access relevant alternatives to compute SIs, it is possible that other cues that make lexical scale members easier to see as relevant alternatives (including other quantifiers) could help children spontaneously generate SIs, even if the stronger scalar itself is not explicitly mentioned.

4.1. Method

4.1.1. Participants

We tested a new group of 60 typically developing 5-year-old children (4;7 – 5;10, M = 5;1) and 24 adult controls, all monolingual speakers of English. None of these participants had taken part in Experiments 1 or 2. The children were recruited from daycare centers in Newark, DE. The adults were college students recruited from the University of Delaware and received course credit for their participation. An additional group of 5 children were tested but excluded from the analysis for failure to follow instructions (n = 3), failure to complete the experiment (n = 1) and experimenter error (n = 1).

4.1.2. Materials and procedure

Participants were equally distributed across two between-subjects conditions: The *All-First* condition was a replication of the *Quantity* condition of Experiment 2. The *None-First* condition was very similar, but the *all* statements in the first block were replaced by *none*-statements. A small change in the scenes for the first block was also made to produce *True-None* trials, where none of the blickets had the item (the *False-None* trials did not require a modification; see Fig. 3 for examples).

4.2. Predictions

An account on which children have problems with spontaneously generating the stronger lexical scale member expects a strong difference between the *All-First* and the *None-First* condition, with only the first one facilitating SI-calculation in children (by directly supplying the hard-to-generate stronger scalemate). Such an account can be found for example in Tieu et al. (2015): "... children have the ability to compute inferences – when the alternatives are explicitly mentioned, either in the discourse context or in the assertion." (ibid., p.25).

However, an account that does not limit children's difficulties to the lexical retrieval of the stronger competitor is more flexible. For instance, it allows for the possibility that, since *none* is a quantifier, it might also help children reason about relevant alternatives (i.e., quantifiers more generally), access the relevant stronger alternative *all*, and detect the underinformativeness of later-occurring weak scalar statements. On this hypothesis, the *None-First* condition could also encourage SI-computation in children, perhaps even to the same degree as the *All-First* condition.

One might assume that the lexical retrieval account could also allow *none* to facilitate SI-calculation by giving access to the scale as a whole. Notice, however, that *none* is actually not a member of the <*some*,... *all*> scale (unlike other quantifiers such as *many* and *most*; Horn, 1972; Levinson, 1983): logical scales are based on entailment, and there is no entailment relationship between *none* and *some* (or *none* and *all* for that matter).

4.3. Coding

The coding scheme was identical to the one used for Experiments 1 and 2.

4.4. Results

Adult performance was consistently very high for all conditions and trial types (see Table 5). Fisher's Exact test analyses on 2 × 2 contingency tables for each trial type revealed no significant difference in the numbers of Passers vs. Failers across conditions (*True-All/None*-trials, $p = 1$; *False-All/None*-trials, $p = 1$; *True-Some*-trials, $p = 1$; *True-and-Infelicitous-Some*-trials, $p = 1$). Adults were overwhelmingly pragmatic in the *True-and-Infelicitous-Some* trials.

Children performed quite well across trial types (Table 5). Fisher's Exact Tests on 2 × 2 contingency tables revealed no significant difference in the numbers of Passers vs. Failers across the two conditions for any of the trial types: *True-All/None*-trials ($p = 0.237$), *False-All/None*-trials ($p = 0.532$), *True-Some-trials* ($p = 1$). Most importantly, in the *True-and-Infelicitous-Some*-trials, children were

Table 5
Participant performance in Experiment 3.

Trial type	Classification	Adults Condition		Children Condition	
		All-First	None-First	All-First	None-First
True-All/True-None	Passers	12	12	30	27
	Failers	0	0	0	3
False-All/False-None	Passers	12	12	22	25
	Failers	0	0	8	5
True-Some	Passers	12	12	29	28
	Failers	0	0	1	2
True-and-Inf-Some	Passers	12	11	23	19
	Failers	0	1	7	11

Table 6
Some/all or *Some/none*-knowers' performance in *True-and-Infelicitous-Some* trials of Experiment 3.

Trial type	Classification	Children Condition	
		All-First	None-First
True-and-Inf-Some	Passers	20	16
	Failers	1	5

highly pragmatic in both the *All-First* and the *None-First* condition, with no difference between conditions (Fisher's Exact test, two-tailed, $p = 0.4$).[4]

As in previous experiments, when asked to justify their rejections of *True-and-Infelicitous-Some* statements, children provided reasonable justifications. They typically either referenced the stronger scalar term ("*All* of them have an X"; 39 out of 42 Passers, or 93%), or mentioned the number of blickets possessing an item ("Because four blickets have an X" 3 out of 42, or 7%). Adults again performed similarly, referencing the stronger scalar term (22 out of 23 Passers, or 96%) or the number of blickets and items (1 Passer, or 4%).

For the same reasons as in the previous experiments, we conducted a second analysis using the semantic trials as controls and excluding children who were Failers in any of semantic trial types (*True-None*, *False-None*, or *True-Some* statements). This resulted in n = 9 children being excluded in the *All-First* condition, and n = 9 children excluded in the *None-First* condition. All of the remaining children can be assumed to have the correct semantics for *some* and either *all* or *none*. A Fisher's Exact test on the 2×2 contingency table in Table 6 revealed no difference between the numbers of Passers vs. Failers for the two different conditions ($p = 0.18$), confirming the results of the first analysis.

Finally, we compared the performance of the children who had a solid grasp of quantifier semantics with that of adults with Fisher's Exact Test on 2×2 contingency tables. No differences were found between age groups in either the *All-First* condition ($p = 1$) or the *None-First* condition ($p = 0.379$).[5]

[4] A binary logistic regression run with Condition (All-First, None-First) as a predictor and children's performance in the *True-and-Infelicitous Some* trials (Passer, Failer) as a binary dependent variable, returned no effect of Condition: *Wald's* χ^2 (1) = 1.254, $p = .263$.

[5] We replicated the *None-First* condition using scenes in which all 4 blickets had the item mentioned in *False-None* trials. The results were very similar. Of the 30 new 5-year-olds who participated, 19 passed and 11 failed the critical *True-and-Infelicitous-Some* trials, a pattern that did not differ from either the original *None-First* ($p = 1$) or the *All-First* condition in Exp.3 ($p = 0.5675$). After exclusions, there were 17 Passers and 3 Failers, a distribution that again was not different from either the original *None-First* ($p = 0.6965$) or the *All-First* condition ($p = 0.3433$) in Exp.3.

4.5. Discussion

In Experiment 3 we compared the effects of explicitly providing the stronger scalar term (*all*) vs. another quantifier (*none*) on young children's success in later deriving SIs from the use of weak scalar statements (e.g., "Some of the Ys have an X"). If activation of the stronger lexical alternative was solely responsible for limitations in children's computation of an SI only the stronger scalar term should facilitate subsequent SI-calculation. However, if recognizing scalar terms as relevant alternatives was also part of the limitations in children's SI computations then either quantifier could encourage later SI-generation as long as the quantifiers encouraged children to consider relevant alternatives to the weak scalar term *some*.

The results support the second possibility: children were overwhelmingly pragmatic in judging *True-and-Infelicitous-Some* statements, even when the stronger scalar term *all* was lexically absent. In fact, after controlling for knowledge of semantics, children in the *None-First* condition were pragmatic at levels comparable to children in the *All-First* condition, and in both cases their performance was adult-like. Thus even when *all* was not explicitly present, children were led by a manipulation that established the relevance of the domain of alternatives (i.e., quantifiers), to spontaneously consider *all* as a relevant alternative and generate a SI.

5. General discussion

5.1. The acquisition of scalar implicatures

In the present series of experiments, we investigated the development of pragmatic inference in children using scalar implicature as a case study. We focused on a factor that has been claimed to bear a major part of the responsibility for children's difficulties with SIs, the accessibility of scalar alternatives (Barner et al., 2011; Chierchia et al., 2001; Gualmini et al., 2001; cf. Bale & Barner, 2013; Papafragou & Skordos, 2016). Our main goal was to uncover the mechanisms whereby the accessibility of the stronger scalar alternative affects pragmatic computations. Specifically, we wanted to adjudicate between accounts that place emphasis on the role of the lexical retrieval of the stronger alternative (either alone or in combination with processing factors involved in maintaining and comparing alternatives in working memory; Chierchia et al., 2001; Gualmini et al., 2001; Tieu et al., 2015) and other accounts that view children's difficulty with SIs as the product of a mechanism that critically also evaluates conversational relevance (Papafragou & Musolino, 2003; Papafragou & Tantalou, 2004; Pouscoulous et al., 2007).

We focused on the quantificational scale <*some*, *all*>. In Experiment 1, we found that the accessibility of the stronger scalar term (*all*) facilitates SI generation from the use of *some* in 5-year-old

children. However, in that experiment, accessible lexical alternatives were also relevant alternatives. In Experiment 2, we independently manipulated whether the stronger lexical alternative could be viewed as relevant or not, while maintaining its lexical accessibility (through externally initiated activation). We found that relevance affects children's use of alternatives during SI generation: children used the explicitly mentioned stronger alternative to successfully generate the corresponding SI only when the alternative was relevant to the dimension that was implicitly set as the basis for evaluating statements within the task. In Experiment 3, we replaced the strong scalar alternative with another quantifier (*none*). We found that, even when we did not provide the stronger scalar explicitly, children were consistently able to generate SIs at adult-like levels based on the relevance of the alternatives.

Together, our results confirm the idea that the accessibility of scalar alternatives affects the calculation of SIs. Crucially, our results offer the first evidence in the literature about the type of mechanism to which children submit accessible alternatives in order to compute SIs. This evidence strongly supports the idea that children's use of alternatives is embedded within a mechanism that takes into account calculations of relevance (as well as quantity). As our studies show, the problem that young children face with SIs might not be solely or primarily sought in their knowledge of or ability to access informationally-asymmetric scalemates but rather in limitations in their ability to quickly and efficiently identify which scalar alternatives are appropriate, i.e., relevant. Thus our findings speak against 'lexical retrieval' accounts that locate children's difficulty with SIs in the lexical accessibility of stronger scalar alternatives (Tieu et al., 2015), or additionally posit processing difficulties with maintaining and comparing different alternatives in working memory (Chierchia et al., 2001; Gualmini et al., 2001).

Our data also point to marked differences in children's and adults' pragmatic ability: as Experiment 2 shows, children, unlike adults, cannot recover conversational relevance flexibly. This conclusion is consistent with studies that have shown that relevance implicatures, a kind of pragmatic inference based predominantly on conversational relevance (e.g., "Mom, can I have some cake? – We are having dinner in a few minutes…"), pose significant difficulties for children up to the age of 6 (Bucciarelli, Colle, & Bara, 2003; de Villiers, de Villiers, Coles-White, & Carpenter, 2009; Loukusa, Leinonen, & Ryder, 2007; Loukusa, Ryder, & Leinonen, 2008; Verbuk & Shultz, 2010). Interestingly, with sufficiently simple tasks, children's performance improves (de Villiers et al., 2009), and even 3-year-olds show some evidence of being able to draw relevance-based inferences, although their performance still falls short of being adult-like (Schulze, Grassman & Tomasello, 2013). The precise circumstances under which young children can assess conversational relevance remain an active topic of investigation (see also next section).

The present data and theorizing allow us to synthesize prior, sometimes conflicting results about children's SI computation into a single coherent picture. To begin with, our data are consistent with previous studies showing that, when children are given a choice between two true statements containing contrastive alternatives (e.g., "Every farmer cleaned a horse or a rabbit" vs. "Every farmer cleaned a horse and a rabbit"), they prefer the stronger, more informative statement, even though they tend to accept underinformative statements when asked to judge them individually (Chierchia et al., 2001; cf. also Ozturk & Papafragou, 2015): in these studies, the stronger alternative is always relevant. Our work is also clearly consistent with findings in which contextual support in the form of background information increased both the salience and the relevance of the stronger alternative and led to higher success with SIs in young children (Foppolo et al., 2012; Papafragou & Musolino, 2003).

More importantly, our approach can accommodate the fact that the explicit mention of a stronger alternative within other experimental tasks did not lead to successful SI generation in children, even though it made the strong scalar alternative salient (Foppolo et al., 2012; Noveck, 2001). Typically, in these prior tasks, there was little to suggest that quantifiers such as *some* and *all* presented across trials should be seen as forming relevant alternatives. In Noveck's (2001) judgment study, for instance, participants were presented with 5 sentences of each of the following types: true *some*-statements ("Some birds live in cages"), true *all*-statements ("All elephants have trunks"), patently false (what Noveck (2001) called "absurd") *some*-statements ("Some stores are made of bubbles"), patently false *all*-statements ("All chairs tell time"), plainly false *all* statements ("All dogs have spots") and true but underinformative *some* statements ("Some giraffes have long necks"). Each statement involved a different state of affairs, thereby weakening the potential of the quantifiers to be used contrastively. Furthermore, the grounds for agreeing or not with the statements were left open (i.e., the relevance of the stronger alternative was not clearly specified): if statements were to be judged on the basis of whether they were true or not, there was no reason to compare them to different alternatives; but if statements were to be judged for felicity, they would have to be implicitly compared to other (relevant) statements the speaker might have uttered. Presumably because of the open-endedness of what was relevant, even adults (who are taken to be fully pragmatically competent) based their responses on logical, not pragmatic, content and agreed with underinformative *some* statements half of the time in this task. In sum, here, as in our own Experiment 2, selective effects of otherwise accessible lexical alternatives on SI derivation can be accounted for by the position that the salience of alternatives is necessary but not sufficient for deriving a SI.

5.2. Beyond alternatives: Pragmatic immaturity vs. pragmatic tolerance

Our data bear on two other major types of accounts that do not place particular emphasis on the role of alternative accessibility for children's apparent difficulty with SIs. The first of these accounts (Noveck, 2001) attributes children's difficulty to their immature pragmatic mechanisms. For instance, Noveck suggests that young children might have difficulty with detecting violations of the maxim of Quantity and that these difficulties are overcome as children's pragmatic mechanisms mature (after the age of 10). There is no evidence in our data that children have any difficulty in detecting Quantity violations, assuming that children have access to relevant alternatives, either when those alternatives were explicitly provided (Experiment 1, Mixed Condition; Experiment 2, Quantity Condition) or when relevant alternatives were indirectly encouraged (Experiment 3). In both of these cases children readily rejected underinformative statements and provided adult-like justifications for their rejections.

The second account claims that children have no difficulty accessing stronger alternatives but nevertheless do not reject weak, underinformative statements such as *Some Xs Y* because they are more pragmatically tolerant than adults (Katsos & Bishop, 2011). As evidence for this hypothesis, Katsos and Bishop show that 5-year-old children are more likely to distinguish between true, false, and true but underinformative scalar statements in a judgment task if given a 3-point Likert scale that uses a small, medium and large strawberry as "rewards" as opposed to a binary response choice: the rationale is that a binary choice leads children to reserve *No* responses only for false, as opposed to true but infelicitous statements but a 3-point scale allows children to make more nuanced responses (i.e., to choose intermediate rewards - the medium strawberry - for true but infelicitous

statements). Our data show that the position that children cannot show their pragmatic competence in binary choice judgment tasks cannot be sustained. Under circumstances where the relevant alternatives are made accessible enough (directly, as in Experiments 1 and 2, or indirectly, as in Experiment 3), children have no problem rejecting true but underinformative statements. More generally, the patterns in the present data cannot be explained by pragmatic tolerance. We believe that the failures observed by Katsos and Bishop (2011) in their binary judgment task were due to the fact that the relevance of the stronger alternative was not made clear before children had to assess true but infelicitous statements (see discussion of similar past results in the previous section). Following the same reasoning, successes in detecting infelicity triggered by the 3-point scale might have been due to the fact that the 3-way distinction suggested to children that the 'goodness' of each statement was gradient and should therefore always be evaluated with respect to some (unspecified) alternative(s).[6]

5.3. Further issues: Relevance and alternatives in pragmatic computation

The present data raise several further issues that remain ripe for future research. As already alluded to, a first issue is how relevance constrains the search for alternatives by both children and adult communicators during the computation of SIs. Cues to relevance are recoverable from a variety of sources, including discourse or information-structure cues, and children's use of such cues appears to be flexible and task dependent (see also de Villiers et al., 2009; Shulze et al., 2013). A critical question is whether children's assessment of relevant alternatives is ultimately constrained by speaker awareness and knowledge state, as should be the case on a fully Gricean rich-computation model where SIs are a sub-type of mental-state inference (see Grice, 1975; Sperber and Wilson, 1986/1995; Carston, 1995; Noveck & Sperber, 2007; cf. Bergen & Grodner, 2012 and Breheny et al., 2013, for discussion). The present data do not establish this conclusion, even though they are compatible with it (cf. Papafragou, Friedberg, & Cohen, 2014)

Relatedly, our studies highlight cases where alternatives are established as part of the process of recovering what is relevant. However, they leave open the possibility that purely linguistic information (e.g., lexical scales, focus etc.) might play an independent role in determining the set of alternatives. A fuller picture of how alternatives are derived during the computation of pragmatic inference needs to move beyond the logical/quantificational expressions considered here to also examine how contextually defined alternatives are accessed and used to compute scalar implicatures (Hirschberg, 1985; see example (5b) in the Introduction; cf. also Papafragou & Tantalou, 2004; Barner et al., 2011; Katsos & Bishop, 2011; Stiller et al., 2015). Some prior studies have shown that logical scales yield harder-to-access alternatives compared to ad hoc, context-based scales (e.g., Barner et al., 2011; cf. Ozturk & Papafragou, 2015). An interesting question is whether these differences between scale types could be reinterpreted in terms of expected relevance along the lines suggested here.

An overarching question in dealing with the above issues is how relevance should be understood in pragmatic computation. One possibility is to pursue a theoretical definition of relevance as the "Question under Discussion" (QUD; Roberts, 1996, 2004; cf. Stalnaker, 1979). According to QUD accounts of pragmatics, discourse is based on conversational goals, foremost among which is an attempt by the communicative partners to discover the state of affairs that obtains with regard to their topic of conversation. In their attempt to do so, communicative partners posit and answer a series of explicit and implicit questions relevant to the aforementioned topic. An utterance is considered relevant to the QUD if it provides a (full or partial) answer to it (see Groenendijk & Stokhof, 1984; van Rooij & Schulz, 2004; Sauerland, 2004, for applications of the QUD model to SIs; and Russell, 2012; Zondervan, 2010, for critical comments). While our results do not necessarily commit one to a very rich pragmatic account, where conversational goals and the QUD are actively shared between interlocutors, this remains a possible interpretation of the findings. Our current experimental set-up might not qualify as a 'true conversation' or 'naturalistic discourse'; however, experimental settings are rich in pragmatic interpretations and create expectations in the minds of participants trying to interpret the experimenter's instructions and stimuli as pieces of ostensive communication.

Another possibility involves defining relevance as a balance between expected cognitive gains and the amount of cognitive effort incurred in computing those gains (Sperber & Wilson, 1986/1995; see also Frank & Goodman, 2012; Russell, 2012, for additional perspectives). It remains an interesting question whether experimental (and more specifically, developmental) data can offer evidence for or against each of these directions.

[6] An alternative, theoretically less interesting possibility is that children might have been unsure about how to evaluate the true but underinformative statements in the context of a Likert scale and thus chose the medium strawberry in those cases.

Acknowledgements

The work of the first author was partially supported by a Dissertation Fellowship as well as a College of Arts & Sciences Dean's Doctoral Student Summer Scholarship while he was a graduate student at the University of Delaware. We would like to thank the Early Learning Center at the University of Delaware and all other preschools that participated in the studies. We also thank the members of the Language & Cognition lab at the University of Delaware for their help with data collection, as well as the audiences at BUCLD 36 and 38; BCCCD 13 and SRCD 2015 for their helpful feedback.

Appendix A. Supplementary material

The raw data for this paper has been archived in the Open Science Framework and can be accessed online at https://osf.io/s32tw/ under a CC BY-NC 4.0 license.

References

Bale, A., & Barner, D. (2013). Grammatical alternatives and pragmatic development. In A. Falaus (Ed.), *Alternatives in semantics* (pp. 238–266). Palgrave: Macmillan.
Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition, 188*, 87–96.
Barner, D., Chow, K., & Yang, S. J. (2009). Finding one's meaning: A test of the relation between quantifiers and integers in language development. *Cognitive Psychology, 58*, 195–219.
Bergen, L., & Grodner, D. J. (2012). Speaker knowledge influences the comprehension of pragmatic inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 1450.
Bott, L., Bailey, T. M., & Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. *Journal of Memory and Language, 66*, 123–142.
Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language, 51*, 437–457.
Braine, M., & Rumain, B. (1981). Children's comprehension of "or": Evidence for a sequence of competencies. *Journal of Experimental Child Psychology, 31*, 46–70.
Breheny, R., Ferguson, H. J., & Katsos, N. (2013). Taking the epistemic step: Toward a model of on-line access to conversational implicatures. *Cognition, 126*, 423–440.
Breheny, R., Katsos, N., & Williams, J. (2006). Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition, 100*, 434–463.
Bucciarelli, M., Colle, L., & Bara, B. G. (2003). How children comprehend speech acts and communicative gestures. *Journal of Pragmatics, 35*, 207–241.
Carston, R. (1995). Quantity maxims and generalized implicature. *Lingua, 96*, 213–244.

Chierchia, G. (2004). Scalar implicatures, polarity phenomena and the syntax/pragmatics interface. In A. Belletti (Ed.), *Structures and beyond* (pp. 39–103). Oxford: Oxford University Press.
Chierchia, G., Crain, S., Guasti, M. T., Gualmini, A., & Meroni, L. (2001). The acquisition of disjunction: Evidence for a grammatical view of scalar implicatures. *Proceedings from the Annual Boston University Conference on Language Development, 25*, 157–168.
Chierchia, G., Fox, D., & Spector, B. (2009). Hurford's constraint and the theory of scalar implicatures. In P. Egré & G. Magri (Eds.), *Presuppositions and implicatures. Proceedings of the MIT-Paris workshop* (pp. 47–62). Cambridge, MA: MIT Working Papers in Linguistics.
de Villiers, P. A., de Villiers, J., Coles-White, D. J., & Carpenter, L. (2009). Acquisition of relevance implicatures in typically-developing children and children with autism. In J. Chandlee, M. Franchini, S. Lord, & G. M. Rheiner (Eds.), *Proceedings of the 33th annual boston university conference on language development* (pp. 121–132). Somerville, MA: Cascadilla Press.
Feeney, A., Scrafton, S., Duckworth, A., & Handley, S. (2004). The story of *some*: Everyday pragmatic inference by children and adults. *Canadian Journal of Experimental Psychology, 58*, 121–132.
Foppolo, F., Guasti, M. T., & Chierchia, G. (2012). Scalar implicatures in child language: Give children a chance. *Language Learning and Development, 8*, 365–394.
Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science, 336*. 998–998.
Geurts, B. (2010). *Quantity implicatures*. Cambridge: Cambridge University Press.
Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.). *Syntax and semantics: Speech acts* (Vol. 3, pp. 41–58). New York: Academic Press.
Groenendijk, J., & Stokhof, M. (1984) Ph.D. Thesis. University of Amsterdam.
Gualmini, A., Crain, S., Meroni, L., Chierchia, G., & Guasti, M. T. (2001). At the semantics/pragmatics interface in child language. In *Proceedings of semantics and linguistic theory XI*. Ithaca, NY: CLC Publications, Department of Linguistics, Cornell University.
Guasti, M. T., Chierchia, G., Crain, S., Foppolo, F., Gualmini, A., & Meroni, L. (2005). Why children and adults sometimes (but not always) compute implicatures. *Language and Cognitive Processes, 20*, 667–696.
Hirschberg, J. (1985). *A theory of scalar implicature* Doctoral diss. University of Pennsylvania.
Horn, L. R. (1972). *On the semantic properties of the logical operators in English* Doctoral diss. UCLA.
Huang, Y. T., & Snedeker, J. (2009a). Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive Psychology, 58*, 376–415.
Huang, Y. T., & Snedeker, J. (2009b). Semantic meaning and pragmatic interpretation in five-year olds: Evidence from real time spoken language comprehension. *Developmental Psychology, 45*, 1723–1739.
Katsos, N., & Bishop, D. V. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition, 120*, 67–81.
Levinson, S. C. (1983). *Pragmatics. Cambridge Textbooks in Linguistics*. Cambridge: Cambridge University Press.
Levinson, S. (2000). *Presumptive meanings*. Cambridge, MA: MIT Press.
Loukusa, S., Leinonen, E., & Ryder, N. (2007). Development of pragmatic language comprehension in Finnish speaking children. *First Language, 27*, 279–296.
Loukusa, S., Ryder, N., & Leinonen, E. (2008). Answering questions and explaining answers: A study of Finnish-speaking children. *Journal of Psycholinguistic Research, 37*, 219–241.
Matsumoto, Y. (1995). The conversational condition on horn scales. *Linguistics and Philosophy, 18*, 21–60.
Noveck, I. (2001). When children are more logical than adults: Experimental investigations of scalar implicatures. *Cognition, 78*, 165–188.
Noveck, I. A., & Sperber, D. (2007). The why and how of experimental pragmatics: The case of 'scalar inferences'. In N. Burton-Roberts (Ed.), *Advances in pragmatics*. Palgrave: Basingstoke.
Ozturk, O., & Papafragou, A. (2015). The acquisition of epistemic modality: From semantic meaning to pragmatic interpretation. *Language Learning and Development, 11*(3), 191–214.
Papafragou, A. (2006). From scalar semantics to implicature: Children's interpretation of aspectuals. *Journal of Child Language, 33*, 721–757.
Papafragou, A., Friedberg, C., & Cohen, M. (2014). Linking conversational inferences to the speaker's knowledge state. In *Talk delivered at the 38th Boston university conference on language development, 7–9 November*.
Papafragou, A., & Musolino, J. (2003). Scalar implicatures: Experiments at the semantics pragmatics interface. *Cognition, 86*, 253–282.
Papafragou, A., & Skordos, D. (in press). Scalar implicature. In J. Lidz, W. Snyder, & J. Pater (Eds.), *The oxford handbook of developmental linguistics*. Oxford: Oxford University Press.
Papafragou, A., & Tantalou, N. (2004). Children's computation of implicatures. *Language Acquisition, 12*, 71–82.
Pouscoulous, N., Noveck, I., Politzer, G., & Bastide, A. (2007). Processing costs and implicature development. *Language Acquisition, 14*, 347–375.
Roberts, C. (1996). Information structure: towards an integrated theory of formal pragmatics. *OSU working papers in linguistics 49*. Papers in Semantics (pp. 91–136).
Roberts, C. (2004). Information structure in discourse. *Semantics and Pragmatics, 5*, 1–69.
Russell, B. (2012). *Probabilistic reasoning and the computation of scalar implicatures* Unpublished doctoral dissertation. Brown University.
Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy, 27*, 367–391.
Schulze, C., Grassmann, S., & Tomasello, M. (2013). 3-year-old children make relevance inferences in indirect verbal communication. *Child development, 84*(6), 2079–2093.
Smith, C. L. (1980). Quantifiers and question answering in young children. *Journal of Experimental Child Psychology, 30*, 191–205.
Sperber, D., & Wilson, D. (1986/1995). *Relevance: Communication and cognition* (2nd ed.) Cambridge, MA: Harvard University Press.
Stalnaker, R. (1979). Assertion. In P. Cole (Ed.), *Syntax and semantics 9* (pp. 315–332). New York: Academic Press.
Stiller, A., Goodman, N. D., & Frank, M. C. (2015). Ad-hoc implicature in preschool children. *Language, Learning and Development, 11*(2), 176–190.
Tieu, L., Romoli, J., Zhou, P., & Crain, S. (2015). Children's knowledge of free choice inferences and scalar implicatures. *Journal of Semantics*, 1–30.
van Rooij, R., & Schulz, K. (2004). Exhaustive interpretation of complex sentences. *Journal of Logic, Language and Information, 13*, 491–519.
Verbuk, A., & Shultz, T. (2010). Acquisition of relevance implicatures: A case against a rationality-based account of conversational implicatures. *Journal of Pragmatics, 42*, 2297–2313.
Zondervan, A. (2010). *Scalar implicatures or focus: An experimental approach* Unpublished doctoral dissertation.Amsterdam: Universiteit Utrecht.