A speed-accuracy trade-off in children’s processing of scalar implicatures

Rose M. Schneider  
rschneid@stanford.edu  
Department of Psychology  
Stanford University

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

Abstract
Scalar implicatures— inferences from a weak description (“I ate some of the cookies”) that a stronger alternative is true (“I didn’t eat all”)— are paradigm cases of pragmatic inference. Children’s trouble with scalar implicatures is thus an important puzzle for theories of pragmatic development, given their communicative competence in other domains. Previous research has suggested that access to alternatives might be key. Here, we explore children’s reaction times in a new paradigm for measuring scalar implicature processing. Alongside failures on scalar implicatures with “some,” we replicate previous reports of failures with “none,” and find evidence of a speed-accuracy trade-off for both quantifiers. Motivated by these findings, we explore the relationship between accuracy and reaction time with a Drift Diffusion Model. We find evidence consistent with the hypothesis that preschoolers lack access to the alternatives for scalar implicature computation, although this set of alternatives may be broader than previously assumed.

Keywords: Pragmatics; development; scalar implicature; diffusion models.

Introduction
Language comprehension in context is an inferential process. Listeners are not limited to interpreting the literal meaning of speakers’ utterances; they can also reason about what the speaker intended, based on alternative utterances. In the case of pragmatic implicatures (Grice, 1975), a speaker employs a weaker literal description to imply that a stronger alternative is true. Adult listeners tend to infer from the statement “I ate some of the cookies” that some, but not all, of the cookies remain. This scalar implicature (SI) relies heavily on a knowledge of the relevant lexical alternatives in the quantifier scale <some, all>. On standard theories, a listener must be able to contrast “some” with the stronger descriptor “all” to compute the implicature (Grice, 1975; Levinson, 2000).

SIs are challenging for children until surprisingly late in development (Noveck, 2001). For example, when judging a scene in which three of three horses have jumped over a fence, five-year-olds are likely to endorse the statement “the horses jumped over the fence” as felicitous, despite the availability of a more informative alternative (“all”; Papafragou & Musolino, 2003). Children do seem to have some knowledge of these scalar terms, however; for example, they differentially reward speakers based on the informativeness of their scalar descriptions (Katsos & Bishop, 2011). Given this early sensitivity, why do children still struggle to compute scalar implicatures until fairly late in development?

One possible cause of children’s failures is that they may not have access to relevant lexical alternatives (Barner & Bachrach, 2010). This idea, which we will refer to as the Alternatives Hypothesis, predicts that if children cannot quickly and reliably bring to mind the relevant alternative quantifiers (e.g., “all” in a situation where they hear “some”) they will be unable to compute the implicature. The alternatives hypothesis makes a number of predictions about children’s abilities in reasoning about quantifiers, some of which have been confirmed empirically. For example, consistent with the idea of inaccessible alternatives, Barner, Brooks, & Bale (2011) showed that four-year-olds could not compute the quantifier expression “only some” (which should force alternatives to be negated semantically, rather than pragmatically). In light of this hypothesis, what are the proper alternatives for SIs?

Here, empirical evidence has been changing rapidly. Although the conventional view on SI is that the primary inferential alternative is “all,” a new body of evidence suggests that more alternatives may be necessary. For example, Degen & Tanenhaus (2015) found that set size can change the felicity of quantifier SIs for adults: “some” is more felicitous when participants could not say “one” or “two.” In a computational reanalysis of Degen & Tanenhaus (2015) and other data, Franke (2014) showed that a high weight on the alternative “none” was critical for fitting these data. And in a recent study with children, Skordos & Papafragou (in press) found that exposing children to either “all” or “none” facilitated computation of subsequent SIs. Taken together, these data suggest that the availability of alternatives—particularly “none”—does affect scalar implicature processing.

This relationship to “none” is unexpected on classic Gricean theories (Grice, 1975; Horn, 1972), where the only alternatives should be those logically entailed by the original message (i.e. “all”). But it is in fact predicted by recent probabilistic models of implicature. Under these models, all the relevant alternatives compete with one another (Franke, 2014; Goodman & Stuhlmuller, 2013). On the other hand, all of the evidence cited above for the claim of “none” as an alternative is relatively indirect, and such a substantial revision to theory requires further evidence.

One other recent developmental study further supports the importance of “none” in SIs and provides the starting point for our current experiment. Horowitz & Frank (2015) designed a referent selection paradigm that could be used across a broad age range (3–5 years) to explore both scalar and ad-hoc (context-dependent) implicatures. In this task, children saw three book covers, each featuring four familiar objects (Figure 1). On target trials, the experimenter described a book using a semantically ambiguous description (e.g., “On the cover of my book, some of the pictures are cats” [scalar] or “On the cover of my book are cats” [ad hoc]). Children succeeded on ad-hoc trials but largely failed to make SIs, sug-
suggesting they had the pragmatic competence necessary to compute the implicature, but failed to do so for scalar descriptors.

Interestingly, in Horowitz & Frank (2015), the same children who failed on SI also failed on unambiguous “none” control trials (e.g., “On the cover of my book, none of the pictures are cats”) — and in several samples, performance was highly correlated between “none” and “some” trials. This result would be predicted if “none” were in fact an inferential alternative to “some.” If children were not computing its semantics appropriately in an online fashion, they would fail in the “some” SI computation as well, leading to a correlation.

One further prediction of the alternatives hypothesis relates to processing time. Perhaps children who have a fully-established quantifier scale — and hence can make correct SIs — take additional time in using this information, due to competition between alternatives. Congruent with this prediction, our intuition in Horowitz & Frank (2015) and Horowitz, Schneider & Frank (in prep.) was that when children made correct SIs they appeared to be taking longer than when they failed. Motivated by this observation, we adapted our SI task for the iPad to collect detailed and accurate developmental reaction time data. Although reaction time measures have been commonplace in studies of adults’ SI processing, they have been almost entirely absent in the developmental literature (with the exception of Huang & Snedeker, 2009, whose data showed little evidence of SI computation).

Thus, in our current study, we explore children’s response latencies in an iPad adaptation of the Horowitz & Frank (2015) SI task. In our analyses, we examine overall accuracy and patterns of performance, as in Horowitz & Frank (2015), and find that children not only struggle in making SIs, but replicate the finding that they have difficulty with “none” until fairly late in development. Congruent with our predictions, in reaction time analyses we find evidence of a speed-accuracy trade-off for both quantifiers, such that children who succeed exhibit longer response latencies. Finally, we use a Drift Diffusion Model (Ratcliff, 1978; Ratcliff & Rouder, 1998) to explore the source of this increased reaction time. Overall, our findings are consistent with a version of the Alternatives Hypothesis under which “none” is an important inferential alternative in SI and its availability causes slower processing times but correct SIs.

Table 1: Age information for all participants.

<table>
<thead>
<tr>
<th>Age group</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>3–3.5 years</td>
<td>24</td>
<td>3.27</td>
<td>3.28</td>
<td>0.15</td>
</tr>
<tr>
<td>3.5–4 years</td>
<td>35</td>
<td>3.78</td>
<td>3.73</td>
<td>0.15</td>
</tr>
<tr>
<td>4–4.5 years</td>
<td>25</td>
<td>4.27</td>
<td>4.28</td>
<td>0.15</td>
</tr>
<tr>
<td>4.5–5 years</td>
<td>30</td>
<td>4.76</td>
<td>4.76</td>
<td>0.15</td>
</tr>
<tr>
<td>5–6.5 years</td>
<td>24</td>
<td>5.55</td>
<td>5.59</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Method

We adapted the scalar implicature paradigm developed by Horowitz & Frank (2015) for the iPad. In addition to capturing reaction time (RT) data, this version included more trials, and standardized prosody across all trials, as well as a randomized design.

Participants

Table 1 shows the breakdown of participant age information. Included are 138 children out of a planned sample of 120 participants, recruited from both a local daycare and children’s museum. 39 additional children were excluded from analysis based on planned exclusion criteria of low English language exposure (≤ 75%) or < 50% of trials completed. Included in our sample were 79 females and 59 males.

Stimuli and design

The general format of the task was identical to Horowitz & Frank (2015), with the exception of added items for additional trials. The study was programmed in HTML, CSS, and JavaScript, and displayed to children on a full-sized iPad. Each trial displayed three book covers, each containing a set of four familiar objects (Figure 1). Each session involved 30 trials, with 10 trials per quantifier (“all”, “some”, and “none”). Each audio clip used the same three initial sentence frames (e.g., “On the cover of my book, some of the pictures . . .”) to emphasize prosody equally across all trials. The average length of each clip (including target item phrase, e.g., “. . . are cats”) was approximately 6s. Quantifier trial order, items (within category), target item, and quantifier were randomized for all participants. There were 270 different target items and audio clips.

Procedure

Sessions took place individually in a small testing room away from the museum floor or the classroom of the daycare. To familiarize children with the iPad, each session began with a

Figure 1: Example trial stimuli used in Horowitz and Frank (2015).
“dot game,” which required them to press dots on the screen as fast as possible. After the dot game, the experimenter introduced them to “Hannah,” a cartoon character who wanted to play a guessing game with her books. The experimenter explained that Hannah would show the child three books, and would give one hint about which book she had in mind, so they had to listen carefully. Children then saw a practice trial with an unambiguous noun referent.

Each trial allowed 2.5s for children to visually inspect the book covers before the prompt played (e.g., “On the cover of my book, none of the pictures are cats.”). Reaction times were measured from the onset of the target word. Children could only make one selection. If a child did not hear Hannah’s prompt, the experimenter repeated it, matching the original prosody. Once children made their selection, a green box appeared around the chosen book. The experiment was self-paced, and children initiated each trial by pressing a button that appeared after they made their selection.

Results
For trials where the child had missed the prompt or was not paying attention, we excluded reaction times (RTs) longer than 15s. After this initial cut, we excluded RTs outside three standard deviations of the log of mean reaction time. This cleaning resulted in RT data loss for 85 trials (2.09%). We observed a fairly wide RT distribution for all trial types. This variability in children’s RTs may have been exaggerated by testing in a museum environment, or by the motoric demands associated with making a response on a tablet (Frank, Sugarman, Horowitz, Lewis, & Yurovsky, 2016). “Success” is defined as choosing the image consistent with the scalar description (e.g., selecting the book with two cats and two birds in response to the prompt “On the cover of my book, some of the pictures are cats.”).

Accuracy
Figure 2 shows children’s mean performance for each trial type, split by age group. For each age group, we saw significantly lower accuracy for the quantifiers “some” and “none” in comparison to “all” (all ps < .01 in two-sample t-tests for each age group). These results generally replicate our previous findings using this paradigm (Horowitz & Frank, 2015); one difference from previous results was in implicature trials. Children aged 3–5 years performed significantly lower on “some” (implicature) trials in this task in comparison to Horowitz et al. (in prep.) (p < .01 for all tests). Thus, while the iPad adaptation was generally successful, implicatures were more difficult, perhaps because of the non-social nature of the iPad interaction.

We next fit a logistic mixed effects model predicting correct response as an interaction of age and trial type, with random effects of trial type and participant. Performance was significantly lower on “some” (β = −6.72, p < .0001) and “none” trials (β = −9.44, p < .0001). There was also a significant interaction between age and trial type on “none” trials (β = 1.52, p = .0005), indicating that children’s performance with this difficult quantifier increased with age. This model also showed that children’s performance showed a trend towards significance for “some” trials (β = 0.76, p = .051).

Figure 3 shows distributions of correct responses for all trial types. Performance on “some” and “none” trials was bimodal (Hartigan’s D = 0.08, p < .0001) and “none” trials (D = 0.11, p < .0001). While children’s average accuracy was low for these quantifiers, there were some children who were correct on the majority of these trials (“Some”: N = 30; “None”: N = 37) and the others were typically incorrect on the majority of trials. Children did not appear to be responding randomly. As in previous work, we found a strong correlation between children’s accuracy on “some” and “none” trials (r = 0.49, p < .0001). Children also exhibited some interesting systematicity in their errors: on incorrect “some” trials they overwhelmingly chose “all,” while on “none” trials they also chose “some” at about half the rate of “all.”

---

3 All mixed effects models were fit in R using the lme4 package. The model specification was: correct ∼ age * trial type + (trial type | subject id).
**Reaction time**

We fit a linear mixed effects model predicting log RT on correct trials as a function of log trial number, the interaction of age and trial type, and random effects of trial type by subject.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>$-0.28$, $p &lt; .00001$</td>
</tr>
</tbody>
</table>

Response times were longer on “none” ($\beta = 0.38$, $p < .0001$) and “some” trials ($\beta = 0.22$, $p < .0001$), and reaction times decreased with age ($\beta = -0.29$, $p < .0001$). There were no significant interactions between age and trial type. The model also showed a main effect of trial number, with reaction times decreasing over the course of the study ($\beta = -0.1$, $p < .0001$).

Examination of the pattern in Figure 4 suggests that accuracy and reaction time may interact, however. In particular, while correct responses on “all” trials appear to be faster than the (few) incorrect responses, the opposite is true for “none” and “some” trials: Errors have faster RTs, potentially indicating a speed-accuracy trade-off. To test for this effect, we fit another mixed effects model, this time including accuracy and its interactions with age and trial type as predictors. This model revealed that correct trials overall had faster RTs ($\beta = -0.16$, $p = .0002$), but that this accuracy term interacted negatively with trial type such that both “none” and “some” trials had slower RTs for correct trials ($\beta = 0.34$, $p < .0001$; $\beta = 0.27$, $p < .0001$). There were no three-way interactions of trial-type and age. This model thus provides evidence of a speed-accuracy trade-off for “some” and “none” trials.

**Drift diffusion models**

Motivated by the evidence of a speed-accuracy trade-off, we further explored the interaction between reaction time and accuracy in more depth using drift diffusion modeling (DDM). Response latencies associated with “some” and “none” indicated that these longer RTs were associated with higher accuracy, but what components of the decision process contribute to this finding? DDM can be used in behavioral tasks to provide a more detailed view of the relationship between accuracy and reaction time (Ratcliff & Rouder, 1998). In DDM, a behavioral response (a correct or incorrect choice) is the result of noisy data accumulation through a diffusion process. Responses have separation boundaries that are dependent on the amount of information needed to initiate a response; *drift rate* formalizes the rate of data accumulation. *Nondecision* is the amount of time between stimuli offset and initiating the diffusion process (i.e., encoding). Finally, different responses may be *biased* in their starting point in the diffusion process. Thus, a DDM can reveal more differences in the SI decision-making process, and provide clues about the causes underlying this speed-accuracy trade-off.

**Developmental analyses**

Although DDMs are traditionally fit to data from two-alternative forced-choice tasks, here we estimate the drift process between a correct and incorrect choice, with two options in each trial being “incorrect,” and only one being consistent with the target noun and quantifier. We estimated parameters for each subject for each trial type using the RWiener package. We then aggregated across subjects to obtain means and confidence intervals for each age group. Figure 5 shows the parameter estimates for each age group, split by trial type. To fit the model, we excluded 85 trials with outlier RTs.

For each parameter estimate, we ran a mixed effects model, predicting parameter value as an interaction of age and trial type. There was no significant effect of trial type in boundary separation, indicating that roughly the same amount of information is needed to make a decision for each quantifier. This should be expected, given our experimental design. For non-decision time, we found a significant main effect of age ($\beta = -0.28$, $p < .00001$), as well as an interaction between age and trial type.

---

4. Model specification: \( \log(\text{reaction time}) \sim \log(\text{trial number}) + \text{age} \times \text{trial type} + (\text{trial type} | \text{subject id}) \). Age was centered for ease of interpretation of coefficients, and we calculated $p$ values via the $t = z$ approximation.

5. The specifications for all parameter models are as follows: Parameter Value $\sim$ age * trial type + (1 | subject ID)
Figure 5: Parameter estimates for drift diffusion model, split by age and trial type. Error bars are 95 percent confidence intervals computed by nonparametric bootstrap.

In our bias estimates, however, we found a significant interaction between accuracy group and trial type on “some” trials (β = -0.18, p = .0013). This interaction suggests that bias (the starting point in the diffusion process) might be an important factor in successfully making a scalar implicature: More successful children were less biased towards incorrect response alternatives, perhaps due to greater knowledge about the quantifier scale.

General Discussion

What makes scalar implicatures using quantifiers so hard for children? The best current hypothesis posits that children do not have access to the appropriate inferential alternatives and hence fail to consider them in their pragmatic computation (Barner & Bachrach, 2010; Barner et al., 2011). But what are those alternatives? Recent work has suggested that the negative alternative “none” may compete with “some” and “all” when making SIs. Although “none” is not typically considered an alternative in Gricean theories (Grice, 1975; Horn, 1972), it nevertheless provides a relevant lexical alternative along the quantifier scale. Our findings here are consistent with this account and provide some additional support. Using a new method, we replicated the pattern found in previous studies that those children who succeed in comprehending the quantifier “none” are also able to make SIs (Horowitz & Frank, 2015; Horowitz et al., in prep.). In addition, our data revealed a speed-accuracy trade-off, such that reaction times in those trials for which children succeeded in making SIs were slower overall.

One interpretation of this speed-accuracy trade-off is that children who have more inferential alternatives accessible to them (e.g. are considering “none,” “some,” and “all” together) are both better at making SIs and slower to make them due to the processing cost of making the inference. This theory makes the prediction that children who have greater quantifier knowledge will be more successful in computing SIs, but will display higher RTs due to competing scalar alternatives. Our data are consistent with this account, which is also supported by an exploratory drift diffusion model analysis. We fit a DDM to our data for children who succeeded in making scalar implicatures versus children who failed. The model suggested that bias in “some” and “none” trials might be a key factor related to success—that is, children who were con-
sidering “some” and “none” responses equally in their decision were more likely to make the SI. Both of these findings are again consistent with the idea that weighing alternatives appropriately in the SI computation is critical to success.

The speed-accuracy patterns we report are correlational, however, and other accounts are consistent with these findings as well. For example, some third factor (say inhibitory control) could underlie the ability to succeed in “some” and “none” trials and also explain why some children are able to inhibit their response long enough to complete the SI computation. Horowitz et al. (in prep.) did not find evidence of correlations between individuals’ SI abilities and their executive function using one popular measure (the dimensional change card sort). Other versions of this account (or other accounts entirely) are still possible, however, and should be explored in future work. Nevertheless, our present work suggests that there is a meaningful relationship between children’s accuracy and processing times in making scalar implicatures.

Acknowledgements

Thanks to Bing Nursery School and the San Jose Children’s Discovery Museum. Thanks also to Veronica Cristiano, Rachel Walker, and Tamara Mekler for their help with data collection, and to Kara Weisman and Ann Nordmeyer for their assistance creating stimuli. This work was supported by NSF BCS #1456077.

References


