A speed-accuracy tradeoff 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 not true ("I didn't eat all")—are a paradigm case of pragmatic inference. Children's trouble with scalar implicatures is thus an important puzzle for theories of the development of pragmatics, given their communicative competence in other domains. Here, we explore children's reaction times in a new paradigm for measuring scalar implicature processing. Alongside falures on scalar implicature with "some," we replicate previous reports of failures on the quantifier "none" and find evidence of a speed-accuracy tradeoff for both "some" and "none." Motivated by these findings, we use a Drift Diffusion Model to explore the relationship between accuracy and reaction time in our task and find evidence consistent with the hypothesis that preschoolers lack access to the relevant alternatives for the scalar implicature computation. The set of relevant alternatives may be broader than has been previously assumed, however.

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

Introduction

Language comprehension in context is an inferetial process. Listeners are not limited to interpreting the literal meaning of speakers' utterances; they can also reason about what the speaker intended, based on other utterances the speaker could have said. 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 enjoyed *some* of my winter break" that some, but not all, of the break was pleasant. 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 challeinging 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 "some of the horses jumped over the fence" as felicitous, despite the presence of a more informative alternative ("all"; Papafragou & Musolino, 2003). Children do seem to have some knowledge of these scalar terms, however; for example, they 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 late in development?

One possibile cause of children's failures is that they might not have access to the relevant lexical alternatives (D. Barner & Bachrach, 2010). This idea, which we will refer to as the *Alternatives Hypothesis*, predicts that if children cannot quickly and reliabily bring to mind the relevant alternative quantifiers (e.g., "all" in a situation where they hear "some") they will be unable to make the implicature computation. 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, David Barner, Brooks, & Bale (2011) showed that four-year-olds could not even resolve the quantifier expression "only some" (which should force alternatives to be negated semantically, rather than pragmatically). But what are the proper alternatives for SIs?

With respect to the proper set of alternatives for SI, the 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—in particular, "none." For example, Degen & Tanenhaus (2015) found that set size changes the felicity of quantifier SIs for adults: "some" is more felicitous when you couldn't say "one" or "two." In a computational reanalysis of these 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 their computation of subsequent SIs.

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 (Goodman & Stuhlmller, 2013, Franke (2014)). 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 standard theory requires further evidence.

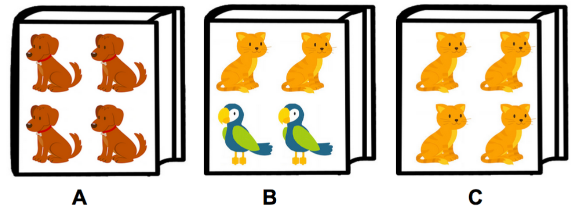

Figure 1: Example trial stimuli used in Horowitz and Frank (2015).

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 succeded on ad-hoc trials but largely failed to make SIs, suggesting they had the pragmatic competence necessary to compute the implicature.

Interestingly, in Horowitz & Frank (2015), the same children who failed on SI also failed on unambiguous "none" control trials—and in several samples, performance was highly correlated between "none" and "some" trials. This result would be predicted if "none" was in fact an inferential alternative. If some children were not computing its semantics appropriately in an online fashion, they would be the children to fail in the SI computation as well.

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 the Horowitz & Frank (2015) study described above and in pilot studies using this same paradigm was that when children made correct SIs they appeared to be taking longer than when they failed. But 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 behavioral response latencies in an iPad adaptation of the Horowitz & Frank (2015) scalar implicature task. In our analyses, we explore overall accuracy and patterns of performance, as in (Horowitz & Frank, 2015), and find that children not only struggle in making a scalar implicature, but replicate the finding that they also grapple with "none" until fairly late in development. Congruent with our predictions, in examining reaction time patterns across all quantifier types, we find evidence of a speed-accuracy tradeoff for both quantifiers. Finally, we use a Drift Diffusion Model 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. We consider this and other alternative explanations in the Discussion.

Age group	N	Mean	Median	SD
3–3.5 years	24	3.27	3.27	0.14
3.5–4 years	35	3.78	3.73	0.15
4–4.5 years	25	4.28	4.28	0.15
4.5–5 years	30	4.76	4.76	0.15
5–6.5 years	24	5.55	5.56	0.36

Table 1: Age information for all participants.

Method

In this study, we adapted the scalar implicature paradigm developed by Horowitz & Frank (2015) for the iPad. In addition to capturing detailed reaction time data, this version included more trials, and standardized prosody across all trials, as well as a completely randomized design.[1]

Participants

Table 1 shows the breakdown of age information for all participants. Included in analyses are 138 children out of a planned sample of 120 participants, recruited from both a local daycare and a local children's museum. We ran 20 additional children, who were excluded from analysis based on planned exclusion criteria of low English language exposure ($\leq 75\%$) or $< 50\%$ of trials completed. Included in our sample were 79 females and 59 males.[2]

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-type ("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…") so that prodosdy was emphasized equally across all trials. The average length of each audio clip (including target item phrase, e.g., "…are cats") was approximately 6s. In our randomization, quantifier triad order, items (within category), target item, and quantifier were randomized for all participants. In all, there were 270 different target items and audio clips.

[1]The full experiment can be viewed online at `https://rosemschneider.github.io/tablet_exp/si_tablet.html` and all of our data, processing, experimental stimuli, and analysis code can be viewed in the version control repository for this paper at: `https://github.com/rosemschneider/SI_tablet`.

[2]Based on Horowitz & Frank (2015), we initially planned to collect data from children 3–5 years. After collecting data from 57 participants, however, we observed significantly lower performance on implicature trials across all age groups, indicating that the iPad adaptation of the scalar implicature task was slightly more challenging for all children, and included an older age group of 24 5–6.5-year-olds.

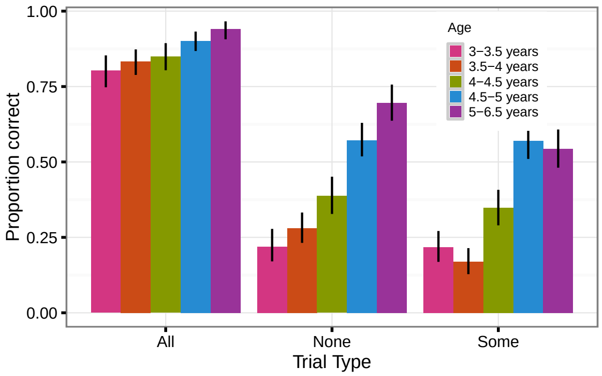

Figure 2: Children's overall accuracy for each quantifier type. Bars show mean performance for each age group. Error bars are 95 percent confidence intervals computed by non-parametric bootstrap.

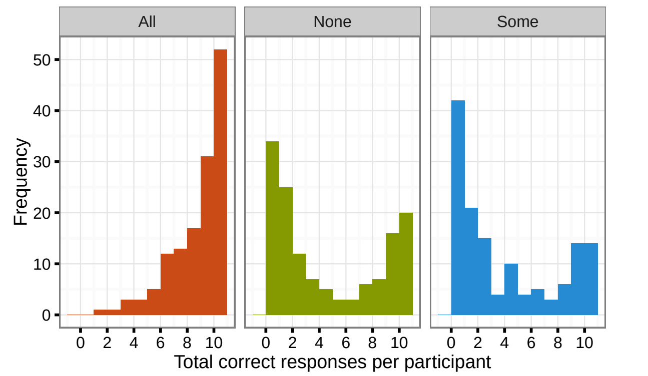

Figure 3: Frequency histogram of correct responses for each trial type, across all participants.

Procedure

Sessions took place individually in a small testing room away from the museum floor or the classrooom of the daycare. To familiarize children with the iPad, each session began with a "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 three book covers, before the experiment played the trial prompt (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 was not paying attention, or if she did not hear Hannah's prompt, the experimenter repeated it, matching the original prosody. Once children correctly 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 had made their selection in the previous trial.

Results

To exclude 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 process resulted in a data loss of 85 trials (2.14%).

Accuracy

Figure 2 shows children's 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), but one difference from the previous results was in implicature trials. Children aged 3–5 years performed significantly lower on "some" (implicature) trials in this task in comparison with data from Horowitz, Schneider, & Frank (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 or the recorded audio stimuli.

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.[3] Performance was significantly lower on "some" ($\beta = -6.98$, $p < .0001$) and "none" trials ($\beta = -9.55$, $p < .0001$). There was also a signficant interaction between age and trial type on "none" trials ($\beta = 1.53$, $p < .0001$), indicating that children's performance with this difficult quantifier increased with age.

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 = 28; "None": N = 36) 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 ($r = 0.49$, $p < .0001$).

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.[4] Reaction times were longer on "none" ($\beta = 0.22$,

[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)`.

[4] Model specification: `log(reaction time) log(trial number) + age * trial type + (trial type | subject`

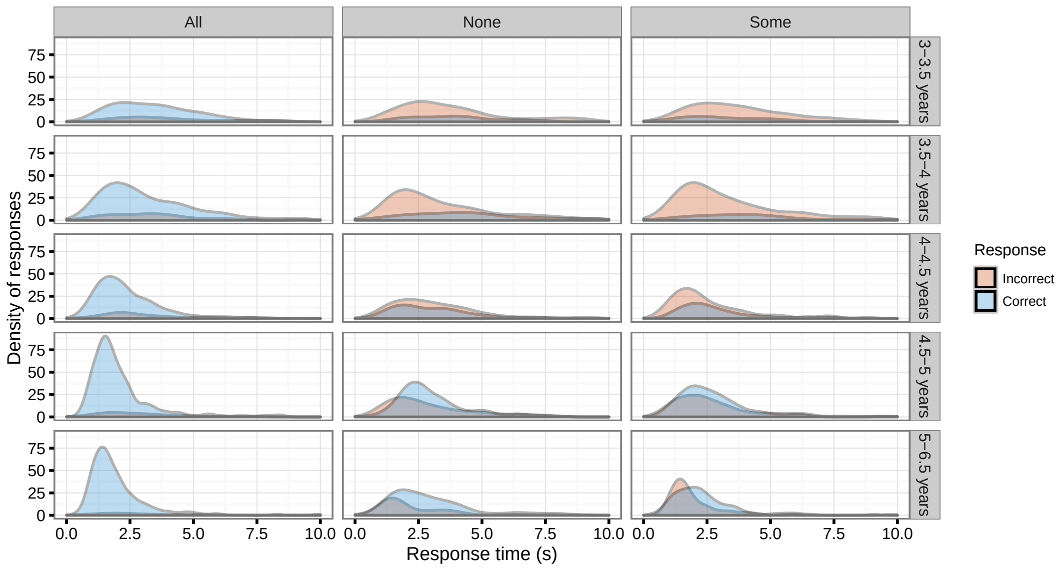

Figure 4: Density plots of reaction times for correct and incorrect responses on each trial type, split by age.

$p < .0001$) and "some" trials ($\beta = 0.1$, $p < .0001$), and reaction times decreased with age ($\beta = -0.1$, $p < .0001$). 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 (β = -0.27, $p < .00001$).

Examination of the pattern in Figure 4 suggests that accuracy and reaction time may be interacting, 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 tradeoff. 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 tradeoff for "some" and "none" trials.

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.[5] Reaction 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 (β = -0.1, $p < .00001$).

Examination of the pattern in Figure 4 suggests that accuracy and reaction time may be interacting, 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 tradeoff. 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 correcvt 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 tradeoff for "some" and "none" trials.

Drift diffusion models

Motivated by the evidence of a speed-accuracy tradeoff we observed, we further explored the interaction between reaction time and accuracy in more depth using drift diffusion modeling. 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 initate a response, and *drift rate* formalizes the rate of data accumulation. *Nondecision* is the amount of time between stimuli offset, and initiating the diffusion process. Finally, different responses may have a *bias*, or different starting point in the diffusion process, dependent on the stimuli.

`id)`. Age was centered for ease of interpretation of coefficients, and we calculated p values via the $t = z$ approximation.

[5] Model specification: `log(reaction time) log(trial number) + age * trial type + (trial type | subject id)`. Age was centered for ease of interpretation of coefficients, and we calculated p values via the $t = z$ approximation.

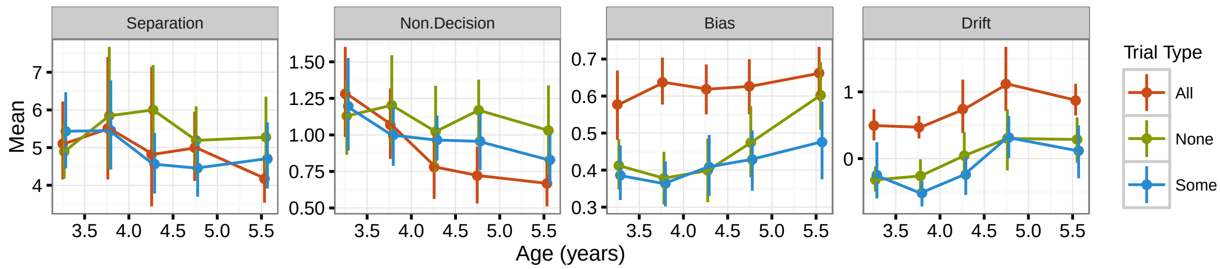

Figure 5: Parameter estimates for drift diffusion model, split by age and trial type. Error bars are 95 percent confidence intervals computed by nonparametric bootstrap.

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.

For each estimate, we ran a mixed effects model, predicting parameter value as an interaction of age and trial type.[6] For boundary separation, there was no significant effect of trial type, indicating that roughly the same amount of information needs to be accumulated to make a decision in each trial type. For nondecision time, we found a significant main effect of age (β = -0.28, $p < .00001$), as well as a interaction between age and "none" trials (β = 0.24, $p = .01$). As expected in drift rate, there was a negative main effect of trial type ("None": β = -1.3, $p = .0185$; "Some": β = -1.18, $p = .03$). Interestingly, for bias there was a significant negative effect of "none" trials (β = -0.5, $p = .0005$), and "some" trials showed a trend towards significance (β = -0.28, $p = .0503$), as well as a significant interaction between age and "none" trials (β = 0.07, $p = .023$). The parameter estimates from our DDM align with the analyses presented above: Older children, who are more familiar with the quantifier scale, are more likely to respond correctly in our scalar implicature task, while younger children's failures appear to be due to a low rate of data accumulation and a high separation boundary.

Exploratory analyses In addition to examining effects of age on the dyiffusion process, we additionally conducted an exploratory analysis, examining differences in the decision-making process for children who consistently made SIs compared with those who did not. We split children by accuracy on scalar implicature trials, and then estimated parameters by accuracy group. High accuracy was defined as an average of 75% or higher performance on scalar implicature trials. Figure 6 shows parameter estimates for each accuracy group, split by trial type.

[6]The specifications for all parameter models are as follows: `Parameter Value age * trial type + (1 | subject ID)`

We again used mixed-effects models to predict DDM coefficients across participants. As in the developmental DDM analysis, there were no significant effects of separation or nondecision. And while drift rates showed a significant effect of accuracy, because we estimated parameters for high- and low-accuracy children separately, these differences are expected. 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.

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 (D. Barner & Bachrach, 2010; David Barner et al., 2011). But what are those alternatives? A variety of recent work has suggested that the negative alternative "none" may compete with "some" and "all." Our findings here are consistent with this account and provide some additional support. We replicated the pattern found in previous studies that those children who succeed in comprehending the quantifier "none" also make SIs (Horowitz & Frank, 2015; Horowitz et al., in prep.). In addition, our data revealed a speed-accuracy tradeoff, such that reaction times in those trials in which children succeeded in making SIs were slower overall.

One interpretation of this speed-accuracy tradeoff 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. Our data are consistent with this account, and we also found some evidence in favor of it from an exploratory drift diffusion model analysis. We fit a DDM to our data for children who succeeded in making scalar implicatures versus children who fail. The model suggested that bias in "some" and "none" trials might be a key factor related to success—that is, children who were considering "some" and "all" responses equally in their decision were more likely to make the SI. This finding again is consis-

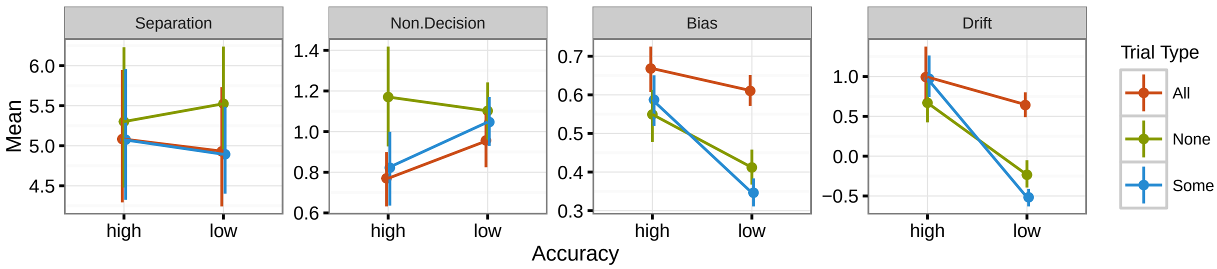

Figure 6: Parameter estimates for drift diffusion model, split by accuracy and trial type. Error bars are 95 percent confidence intervals computed by nonparametric bootstrap.

tent 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 them as well. For example, some third factor (say inhibitory control) could underly 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. Nevertheless, our work here 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.

References

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

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

Degen, J., & Tanenhaus, M. K. (2015). Processing scalar implicature: A constraint-based approach. *Cognitive Science*, *39*(4), 667–710.

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

Goodman, N. D., & Stuhlmller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*(1), 173–184.

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

Horn, L. R. (1972). *On the semantic properties of logical operators.* (PhD thesis). University of California, Los Angeles.

Horowitz, A., & Frank, M. C. (2015). Sources of developmental change in pragmatic inferences about scalar terms. In *Proceedings of the 37th annual conference of the cognitive science society.*

Horowitz, A., Schneider, R. M., & Frank, M. C. (in prep.). The trouble with quantifiers: Children's difficulties with "some" and "none".

Huang, Y. T., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. *Cognitive Psychology*, *58*(3), 376–415.

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

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

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

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

Ratcliff, R., & Rouder, J. (1998). Modeling response times for two-choice decisions. *Psychologial Science*, *9*(5), 347–356.

Skordos, D., & Papafragou, A. (in press). Children's derivation of scalar implicatures: Alternatives and relevance. *Cognition*.