The Effect of Experimental Paradigms on Scalar Implicature Estimation

Background: An intriguing feature of human language is the ability to enrich the literal meanings of utterances with pragmatic implicatures (Grice, 1975; Gazdar 1980; Horn 1972; Levinson 2000; Chierchia 2004). Experimental research on the processing and acquisition of Scalar Implicatures (SIs) relies on behavioral tasks that measure the rate at which SIs are computed within an experimental paradigm. Two paradigms have dominated the experimental pragmatics literature: the *Truth Value Judgment Task* (TVJT) (Gordon, 1998) and the *Picture Selection Task* (PST) (Gerken & Shady, 1998). Yet, the effects of task choice on implicature rate has remained underexplored. Here we report the results of three studies testing participants in the TVJT, PST and a variant of the PST called the Hidden Card Task (HCT) using three different linguistic scales in English: “ad hoc”, “or-and”, and “some-all”.

Methods: In Exp.1, participants responded to both TVJT and PST trials in a single Qualtrics survey. In TVJT trials, participants saw a sentence and a card with animal pictures. They were asked to judge the sentence as true or false. In PST trials, participants saw a sentence and two cards. They were instructed to choose the card that best matched the sentence. In TVJT critical trials (Fig.1a), the description was logically true but pragmatically infelicitous. A “false” judgments counted as evidence for SI computation. In the critical PST trials (Fig.1b), the sentence was logically compatible with both cards, but the implicature of the sentence only matched one card, and thus, choosing that card counted as evidence for implicature computation. To make sure that the within-subjects design did not affect the findings, Exp.2 replicated Exp.1 with a between-subjects (TVJT vs. PST) design. Exp.3 examined a variant of the Picture Selection Task called the Hidden Card Task (HCT) which is being increasingly used in the context of priming research (e.g. Bott & Chemla, 2016). The stimuli used in Exp.3 were adopted from the same inventory of stimuli for the PST in Exp.1 and 2 with an important modification: one card in the stimuli was replaced by a “Better Picture” card. For the critical trials, the “Better Picture” card always replaced the card that matched the implicature of the sentence, while for the control conditions, the “Better Picture” card randomly replaced one of the two cards in the trial (Fig.1c). Each experiment had 18 critical trials and approximately 30 to 40 control trials per task. We recruited 50 participants for each experiment.

Results: For all three experiments, the probability of computing SIs was modelled as a function of task type, scale (“some-all”, “or-and”, and “ad hoc”) along with their interactions using logistic mixed-effects models (Bürkner, 2017). We found main effects of task type, scale and their interactions on the estimated rate of SI computation in both Exp.1 (see Fig.2) and Experiments 2-3 (see Fig.3). Compared with the baseline “or-and” trials, participants in PST computed more SIs in “some-all” trials as well as “ad hoc” trials. For the “or-and” trials, the rate of computing SIs in PST (baseline) was the same as that in TVJT (β = 2.50, CI = [-4.17, 9.69]) and HCT (β = 0.05, CI = [-6.12, 6.37]); however, for the “some-all” trials and “ad hoc” trials, the rates of computing SIs were significantly decreased in the TVJT and HCT as compared with PST.

Conclusions: We found that the estimated rate of SIs is significantly affected by the choice of experimental task and lexical scale. For “ad hoc” and “some-all” scales, TVJT and HCT reported a lower implicature rate than PST. There was no difference in implicature rates for the “or-and” scale across the three tasks. These findings suggest that TVJT and HCT can potentially underestimate participants’ pragmatic abilities, which is central to debates in children’s pragmatic development. They also highlight the special status of exclusivity implications and the possibility that they are fundamentally different from (other) SIs. Finally, our studies stress the need for a more careful attention to the pragmatics of experimental tasks themselves and how they affect participants’ linguistic behavior.

References: Bott, L., & Chemla, E. (2016). Shared and distinct mechanisms in deriving linguistic enrichment. *Journal of Memory and Language*, 91, 117–140. **Bürkner, P. C. (2017).** brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, 80(1), 1–28. **Chierchia, G. (2004).** Scalar Implicatures, Polarity Phenomena, and the Syntax/Pragmatics Interface, in A. Belletti (ed.), *Structures and Beyond: The Cartography of Syntactic Structures*, Volume 3. OUP, 39-103. **Gazdar, G. (1980)**. Pragmatics and logical form. *Journal of Pragmatics*, 4(1), 1-13. **Gerken, L., & Shady, M. E. (1998)**. The picture selection task. In D. McDaniel, C. McKee, & H. S. Cairns (Eds.), *Methods for Assessing Children's Syntax.* The MIT Press. 125–145. **Gordon, P. (1998)**. The truth-value judgment task. In D. McDaniel, C. McKee, & H. S. Cairns (Eds.), *Methods for Assessing Children's Syntax*, The MIT Press. 211-231. **Grice, H. P. (1975)**. Logic and conversation. In P. Cole, & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts.* Academic Press. 41–58. **Horn, L. R. (1972).** *On the semantic properties of logical operators in English*. UCLA Dissertation. **Levinson, S. C. (2000).** *Presumptive meanings: The theory of generalized conversational implicature*. MIT press.

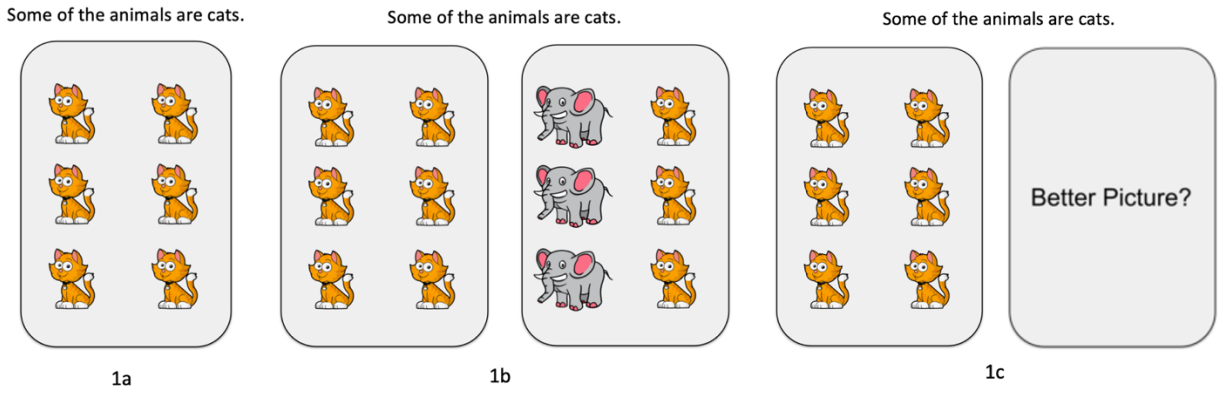

Fig.1 An example of a critical item in TVJT (1a), PST (1b), and HCT (1c). This example concerns the "some-all" scale, while other experimental items may use the "or-and" scale or the "ad hoc" scale. In addition to the images of cats and elephants, images of dogs were also used in the design of the cards. The position of the two cards in PST and HCT was randomized in the experiment.

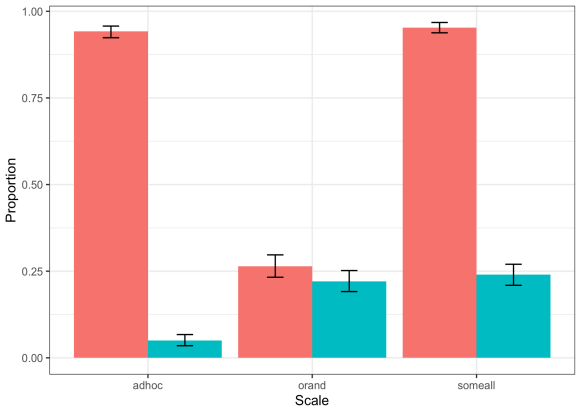

Fig.2: Rate of SI computation estimated by TVJT and PST in Experiment 1. The y-axis shows the percentage of deriving SI for a given scale ("ad hoc" vs "or-and" vs "some-all") in each task (TVJT vs PST), with zero meaning zero percent and one meaning 100 percent. Confidence intervals were computed using bootstrapping methods.

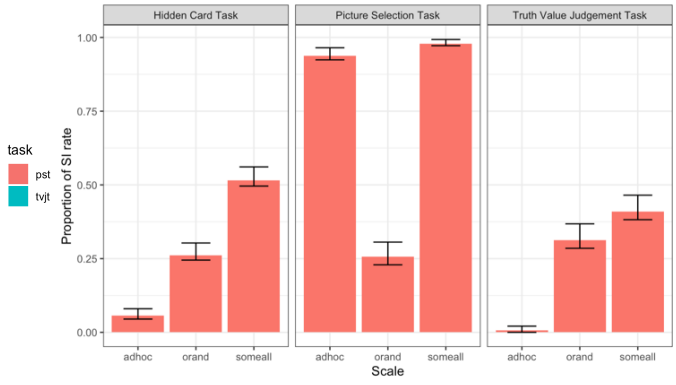

Fig.3: Rate of SI computation estimated by HCT, PST and TVJT in Experiment 2 and 3. The y-axis shows the percentage of deriving SI for a given scale ("ad hoc" vs "or-and" vs "some-all") in each task (HCT vs PST vs TVJT), with zero meaning zero percent and one meaning 100 percent. Confidence intervals were computed using bootstrapping methods.