Experimental Paradigms on Scalar Implicature Estimation

Zhuang Qiu, Casey D. Felton, Zachary N. Houghton & Masoud Jasbi*

Abstract. Experimental research on the processing of Scalar Implicatures (SIs) relies on behavioral tasks that purport to measure the rate at which scalar implicatures are computed within an experimental paradigm. Two paradigms, the Truth Value Judgment Task (TVJT) (Gordon, 1998, Crain & Thornton, 2000) and the Picture Selection Task (PST) (Gerken & Shady, 1998) have dominated the experimental pragmatics literature; yet the effects of task choice on implicature rate have remained underexplored. Here we report the results of three studies testing participants in a TVJT and a PST using three different linguistic scales in English: "or-and", "some-all", and "ad-hoc". We varied the task (TVJT vs. PST) within subjects in the first experiment and between subjects in the second. The third experiment examined a variant of the PST called the Hidden Card Task (HCT) which is increasingly used in the context of priming research (Bott & Chemla, 2016). We found that the estimated rate of scalar implicature computation varied noticeably between different tasks as well as scales. This suggests that the experimental paradigm itself has a significant impact on our estimates of the implicature rate for a given linguistic scale, and thus, researchers studying scalar implicatures need to carefully consider the pragmatics of the task itself when designing experimental studies and interpreting their results.

Keywords. scalar implicature; implicature computation; truth value judgment task; picture selection task; experimental pragmatics

1. Introduction. An intriguing feature of human language is the ability to enrich the literal meanings of utterances with pragmatic implicatures (Grice, 1975, Horn, 1972, Gazdar, 1979, Hirschberg, 1985, Levinson, 1983, 2000), as shown in the following examples:

(1) a. "Some students were late today."
Implicature: Not all students were late today.
b. "Sam had a hot dog or a hamburger for lunch."
Implicature: Sam did not have a hot dog and a hamburger for lunch.
c. "Alice bought a shirt from the store."
Implicature: From that store, Alice only bought a shirt but nothing else.

According to standard accounts (Horn, 1972, Gazdar, 1979, Levinson, 1983), word pairs such as "some-all" and "or-and" form a scale of increasing informativeness with a unidirectional entailment relation. If all students were late today, it must be the case that some students were late, but not vice versa. Therefore, the quantifier "all" is more informative than "some". Similarly, "and" is argued to be more informative than "or" because if Sam had a hot dog and a hamburger for lunch, it must be the case that he had a hot dog or a hamburger for lunch, but not vice versa. The less informative item on the scale is semantically compatible with the cases in which the more informative item holds true; however, the assertion of the lower item implies the negation of the higher item. The computation of such implicature is believed to be governed by general principles of conversation and involves reasoning about the possible alternatives that the

* Zhuang Qiu (zhuangqiu@cityu.edu.mo), City University of Macau; Casey D. Felton (cdfelton@ucdavis.edu), Zachary N. Houghton (znhoughton@ucdavis.edu), and Masoud Jasbi (jasbi@ucdavis.edu), University of California, Davis.

speaker could have said (Grice, 1975, 1978). For example, interlocutors are expected to make their words as informative as required while being truthful at the same time. Abiding by these principles, when hearing the utterance "Some students were late today", the listener reasons that the speaker did not use a more informative alternative "All students were late today" because that alternative is not true. This gives rise to the scalar implicature (SI) that the use of "some" implies the negation of "all" (see Gamut, 1991: 207 for a similar analysis of the "or-and" scale).

The same reasoning process also underlies (1c), which is argued to involve the scale of a set-subset relationship (Hirschberg, 1985:109). In this case, the item "shirt" is a subset of all the items that can be purchased from that store in that conversational context. This type of scale is called an "ad hoc" scale because the implicature is derived on an ad hoc basis depending on the context (Bott & Chemla, 2016). More recent theoretical frameworks either attribute SIs to syntactic operations at the word level (Chierchia, Fox, & Spector 2012, Chierchia, 2013) or highlight conversational contexts rather than lexical scales (Sperber & Wilson, 1995; Degen & Tanenhaus, 2015, 2019). These accounts differ with regard to the mechanisms that derive SIs. However, all of these accounts acknowledge the distinction between upper-bounded inferences (e.g. "some" as "some but not all") and more literal, lower-bounded interpretations (e.g. "some" as "at least one") in expressions containing scalar items as shown in (1a-b).

There is an increasing number of empirical studies on the processing and acquisition of SIs in the relatively new field of experimental pragmatics. Such studies rely on behavioral tasks that aim to measure the rate at which the upper-bounded inferences are computed within an experimental paradigm, among which the Truth Value Judgment Task (TVJT) (Gordon, 1998, Crain & Thornton, 2000) and the Picture Selection Task (PST) (Gerken & Shady, 1998) have dominated the experimental pragmatics literature (as for other widely cited paradigms, see Huang & Snedeker, 2009, 2011, Grodner et al., 2010 for the visual world paradigm, also see Degen & Tanenhaus, 2015 for the gumball paradigm).

In a typical TVJT study, participants are required to judge whether a sentence is true or false based on some background information (the world knowledge shared across the participants or other information provided prior to the target sentence). In the critical trials, the sentence to be judged is pragmatically infelicitous given the background information, but remains logically true (e.g., "Some elephants have trunks"). Therefore, if a participant responds with "false", the experimenter concludes that they have computed an implicature, but not if they respond with "true". In a PST, participants are required to select a picture that best matches a given sentence from a set of pictures. In the critical trials, the sentence is logically compatible with more than one picture, but the implicature of the sentence only matches one picture. For example, the sentence "there is a cat in the picture" is logically compatible with 1) a picture with only a cat and 2) a picture with a cat as well as a dog, but the pragmatic interpretation of the sentence is only compatible with the first picture. Therefore, if a participant selects the first picture, the experimenter concludes that they have computed an implicature.

Recent priming research on scalar implicature (Rees, Carter & Bott, 2023, Rees & Bott, 2018, Bott & Chemla, 2016) adopted a modified version of the PST in which the pragmatically felicitous card (e.g. the card with only a cat in the previous case) is replaced by a card with the text "Better Picture?" on it, and participants were instructed to select the "Better Picture" card if they feel the content of the other card is not optimally described by the given sentence. In this case, the card whose content is fully visible to the participants is the one which is logically compatible with the given sentence but pragmatically infelicitous. Participants' choice between the "Better Picture" card and the fully visible card is interpreted as the presence or absence of a sca-

lar implicature. This version of the PST is referred to as the Hidden Card Task (HCT) since the hypothetical content of the "Better Picture" card is hidden from participants.

Both the TVJT and the PST have been adopted as a measure of SI computation to study topics including but not limited to (1) the debate over the psychological nature of SIs as either default and automatic or as secondary and effortful (Noveck & Posada, 2003, Bott & Noveck, 2004, De Neys & Schaeken, 2007), (2) factors that affect whether a SI is computed in a communicative context (Geurts & Pouscoulous 2008, 2009; Chemla & Spector, 2011, Potts et al., 2015), (3) variations in SI computation across different lexical scales (Doran et al., 2012, van Tiel et al., 2016), and (4) the development of pragmatic knowledge in the first and second language context (Papafragou & Musolino, 2003, Katsos & Bishop, 2011, Horowitz, Schneider & Frank, 2018, Slabakova, 2010, Feng & Cho, 2019).

There seems to be a tacit assumption that different versions of the TVJT and the PST paradigm are variations of objective measures of SI computation, a construct whose psychological nature and structural properties are still debated. This assumption is problematic as there has been increasing concern that the experimental paradigm itself has a significant impact on the estimates of the implicature rate. Papafragou and Musolino (2003) showed that children in general did not judge the underinformative descriptions as "False" in a TVJT task, though they knew that those descriptions were not optimal. Katsos and Bishop (2011) argued that both adults and children are more tolerant to pragmatic infelicity than logical violations, and this explains why underinformative statements tend not to be judged as "False". Crucially, the tolerance to pragmatic infelicity is not equal to the absence of pragmatic inferences, but these two aspects tend to be confused in a binary TVJT task. Thus, Katsos and Bishop included a PST in their study as a complement to TVJT. Since in a PST, participants were asked to select the most felicitous interpretation rather than judge the absolute truth of an interpretation, the "pragmatic tolerance cannot cloud the interpretation of the participants' performance" (2011:14).

Katsos and Bishop brought to the foreground questions such as whether the observed SI rate is contingent on the task choice between TVJT and PST, and if so, how these two paradigms differ in measuring SI computation. However, there has been limited research that compares TVJT and PST in controlled settings using the same set of experimental items with the same group of participants. We conducted three experiments to explore the effect of task variation on the estimated rate of scalar implicature computation. Specifically, we compared the TVJT, PST, and HCT regarding the estimated rate of scalar implicature computation, focusing on the following research questions: 1. When a TVJT, PST, and HCT consisting of theoretically parallel items are administered to the same group of participants, will the observed SI rate differ depending on the task type? 2. How much variation is there in the observed SI computation across different scale types? Is such variation modulated by the task type? 3. How reliable are the TVJT, PST, and HCT as measures of SI computation?

The first two experiments followed a two by three design in which the task type (TVJT vs. PST) and scale type ("some-all", "or-and" and "ad-hoc") were independent variables and the answers elicited were the dependent variable. In the first experiment, both task type and the scale type were manipulated within participants, while the second experiment treated task type as a between participants variable. The third experiment adopted the HCT with experimental items paralleling those of the first two experiments. We found that the reliability of all three tasks are high, but the estimated rate of scalar implicature computation varied noticeably between tasks and scales.

2. Experiment 1. The goal of this experiment was to test the same participants in both the TVJT and the PST tasks. We used three different scales: "or-and", "some-all", and ad hoc. The within-participant design of this study allowed us to see how much the same individual can vary in their responses to each task depending on the linguistic scale.

2.1. METHODS. Fifty participants were recruited from the online crowdsourcing platform Prolific. They were instructed to answer a series of questions corresponding to both a TVJT and a PST in a single Qualtrics survey. In each TVJT trial, participants were presented with a card that had one or more animal images on it and a sentence describing the content of the card. They were instructed to rate the sentence as either true or false. In the critical trials (Figure 1a), the description was logically true but pragmatically infelicitous, and participants' judgment was coded as whether or not a SI was computed. In each PST trial, participants were presented with two cards and a sentence describing the contents of at least one card. Participants were instructed to choose the card that best matched the sentence. In the critical PST trials (Figure 1b), the sentence was logically compatible with both cards, but the implicature of the sentence only matched one card, and thus, participants' preference of one card over the other was coded as whether or not a SI was computed.

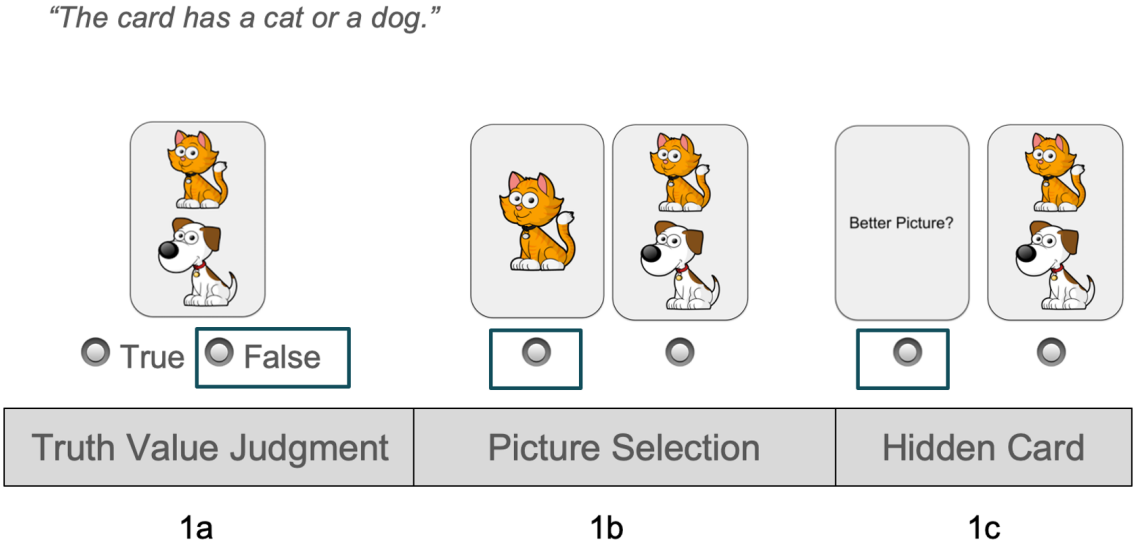

Figure 1: An example of a critical item in TVJT (1a), PST (1b), and HCT (1c). The response showing the computation of SI for each task is marked by the rectangle. This example concerns the "or-and" scale, while other experimental items may use the "some-all" or the "ad-hoc" scale. In addition to the images of cats and dogs, images of elephants were also used in the design of the cards. The position of the two cards in (1b) and (1c) was randomized in the experiment. Task types "TVJT" and "PST" pertain to Experiment 1 and 2, while "HCT" is for Experiment 3.

Critical trials in the TVJT paralleled those in the PST as they used the same picture and the same stimuli sentence. Stimuli sentences came from three different scales to generate SIs: the "some-all" scale, the "or-and" scale, and the "ad-hoc" scale (Bott and Chemla, 2016). In total, there were nine critical trials for each task, and each critical trial appeared twice in the experiment. In addition to the critical trials, 81 control trials were created in the form of either TVJT or PST to check participants' engagement in the task or to provide information about participants' baseline preference (Table 1). In the experiment, critical trials and a subset of control trials (40

out of 81) were blocked, and trials in each block were randomly presented to participants. The control block was always presented after the experimental block, in order to make sure scalar inferences are not affected by the lexical choices in the control trials.

2.2. ANALYSIS. Of the 50 participants recruited, seven were excluded for low accuracy on the attention checking control items, leaving 43 participants in the final analysis. For the purpose of this study, we only analyzed the critical (experimental) trials in the TVJT and the PST.

Condition	Sentence	Task Type	Card (s)	Function
Critical (Experimental)	Some of the animals are cats.	TVJT	6 cats	Measure the computation of SIs
		PST	6 cats AND 3 cats + 3 elephants	
		HCT	6 cats AND "Better Picture?"	
literal maximal (Control)	All of the animals are cats.	TVJT	6 cats OR 3 cats + 3 elephants	Attention checking
		PST	6 cats AND 3 cats + 3 elephants	
		HCT	6 cats AND "Better Picture?"	
literal implicature (Control)	Some but not all of the animals are cats.	TVJT	6 cats OR 3 cats + 3 elephants	Attention checking
		PST	6 cats AND 3 cats + 3 elephants	
		HCT	6 cats AND "Better Picture?"	
literal nonimplicature (Control)	Some and maybe all of the animals are cats.	TVJT	6 cats OR 3 cats + 3 elephants	Measure the baseline preference
		PST	6 cats AND 3 cats + 3 elephants	
		HCT	6 cats AND "Better Picture?"	
Pragmatic implicature (Control)	Some of the animals are cats.	TVJT	3 cats + 3 elephants	Attention checking

Table 1: Experimental manipulations using stimuli from the "some-all" scale as an example. Task types "TVJT" and "PST" pertain to Experiment 1 and 2, while "HCT" is for Experiment 3.

We constructed a Bayesian logistic generalized linear model (Bürkner, 2017) to explore how task variation influences SI computation. The probability of computing SIs was modeled as a function of task type (TVJT vs PST), scale ("some-all", "or-and", and "ad hoc") and their interactions. The model also included by-subject random intercepts, scale-by-subject slopes, task-by-subject slopes, and slopes for the interaction of scale and task by subject. The predictors were dummy coded. Since each critical item appeared twice in the experiment, we also constructed Bayesian logistic generalized linear models to check if the SI rate differed depending on whether the participants encountered the item for the first or the second time. We separated TVJT trials from PST trials, and for each of the tasks, the probability of computing SIs was modeled as a function of trial iteration (first iteration or the second iteration). Maximal random effects structures were constructed including subject and item intercepts and slopes (Barr et al., 2013). We then compared the composite reliability of TVJT and PST based on the coefficient alpha calculated following the Kuder-Richardson Formula 20 (Cronbach, 1951, Kuder & Richardson, 1937).

2.3. RESULTS. We found main effects of task type, scale, and their interactions on the estimated rate of SI computation (Figure 2). Compared with the baseline "or-and" trials (in PST) participants computed more SIs in "some-all" trials (beta = 17.82, CI = [11.34, 27.79]) and "ad hoc" trials (beta = 14.12, CI = [9.11, 21.53]). For the "or-and" trials, the rate of computing SIs in PST (baseline) was the same as that in TVJT (beta = -1.73, CI = [-4.82, 0.98]); however, for the "some-all" trials and "ad hoc" trials, the rates of computing SIs were significantly decreased in the TVJT (beta = -19, CI = [-31.66, -11.26]; beta = -20.84, CI = [-34.98, -12.46], respectively). Moreover, there was no effect of item iteration on the probability of computing SIs for both

TVJT (beta = 0.15, CI = [-0.91, 1.07]) and PST (beta = 0.42, CI = [-0.38, 1.34]). Participants provided the same answer to the same question regardless of whether they saw it the first time or the second time. The coefficient alpha for the 18 TVJT and 18 PST critical items was 0.92 and 0.83, respectively. The observed differences in SI rates were arguably attributable to the nature of TVJT and PST. However, it is possible that the inclusion of both tasks in a within-participant design may prompt different behaviors contingent on the task type. To rule out the possibility that the effects we found were artifacts of the design, we conducted a follow-up study in which task variation was manipulated between participants.

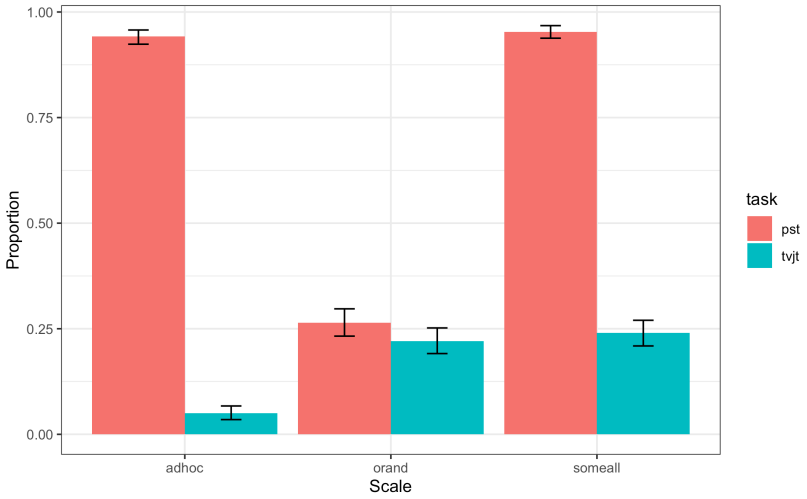

Figure 2: Rate of SI computation estimated by TVJT and PST in Experiment 1. The y-axis shows the percentage of deriving SI for a given scale ("ad hoc" vs "or-and" vs "some-all") in each task (TVJT vs PST), with zero meaning zero percent and one meaning 100 percent. Confidence intervals were computed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

3. Experiment 2. In Experiment 1, we observed varying estimated implicature rate contingent on the task (TVJT vs. PST) and the lexical scale ("some-all", "or-and", and ad hoc). However, it is possible that participants' different responses to different tasks were an artifact of the within-subjects design of the study. Participants saw truth judgment questions and picture selection questions in a random order so they may have decided to treat them differently. In this second study, we ran the tasks between-subjects to address this possible confound.

3.1. METHODS. Experiment 2 adopted the same set of stimuli used in Experiment 1, but the TVJT items and the PST items were a between-subject manipulation rather than a within-subject manipulation. 50 participants were recruited from the online crowdsourcing platform Prolific and were randomly assigned to perform either the TVJT task or PST task. Of the 50 participants recruited, two were excluded for low accuracy on the attention checking control items, leaving 48 participants in the final analysis (24 in TVJT group and 24 in PST group). We combined the Truth Value Judgment Task data with the Picture Selection Task data and adopted a similar set of analyses as in Experiment 1, with minor changes to the random effect structure necessitated by the between-subjects design. Since the task difference was manipulated between participants, the random effect structure of the model only included by-subject random intercepts and scale-by-subject slopes.

3.2. RESULTS. All the main effects and the interaction we observed in Experiment 1 were replicated in Experiment 2 (Figure 3). In the PST task, the "some-all" trials (beta = 28.07, CI = [14.00, 56.33]) and "ad hoc" trials (beta = 26.88, CI = [12.22, 56.56]) received noticeably more pragmatic interpretation than the baseline "or-and" trials. While no statistically meaningful difference in SI rate was observed between TVJT and PST in the "or-and" trials (beta = 1.85, CI = [-5.47, 9.92]), for the other two scales, participants showed noticeably less SI computation in TVJT compared with PST (beta = -28.83, CI = [-61.96, -10.55] for the "some-all" trials; beta = -52.16, CI = [-112.47, -23.12] for the "ad hoc" trials). Moreover, there was no effect of item iteration on the probability of computing SIs for both TVJT (beta = -0.47, CI = [-2.03, 0.63]) and PST (beta = 1.54, CI = [-0.88, 5.98]), the same as what we found in Experiment 1. The coefficient alpha for the 18 TVJT and 18 PST critical items was 0.91 and 0.86, respectively.

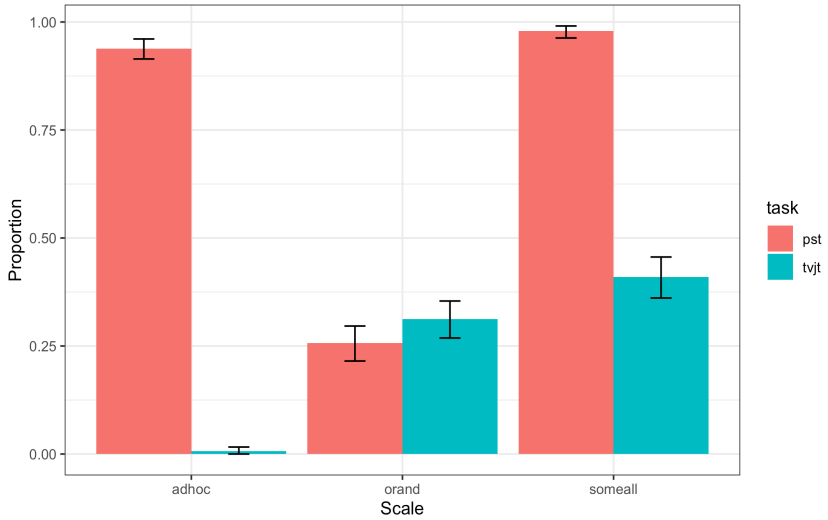

Figure 3: Rate of SI computation estimated by TVJT and PST in Experiment 2. The y-axis shows the percentage of deriving SI for a given scale ("ad hoc" vs "or-and" vs "some-all") in each task (TVJT vs PST), with zero meaning zero percent and one meaning 100 percent. Confidence intervals were computed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

4. Experiment 3. We tested participants on the HCT and compared our results with the findings of the TVJT and PST in Experiment 2.

4.1. METHODS. We recruited 50 participants from the online crowdsourcing platform Prolific. Those participants were instructed to perform a HCT in a single Qualtrics survey. In each trial, participants were presented with two cards and a sentence potentially describing the content of one card. Among the two cards, only one of them had contents visible to the participants, while the content of the other card was covered. Participants were instructed to choose the card that best matched the given sentence. In the critical trials (Figure 1c), the sentence was logically compatible with the visible card, but the implicature of the sentence did not match the visible card, and thus, participants' preference to the visible card was interpreted as the absence of scalar implicature computation, but their preference to the hidden card was coded as scalar implicature computation. The stimuli used in this experiment were adopted from the same inventory of stimuli for the PST in the previous experiments with an important modification: one card in the stimuli was replaced by the "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. An example of the experimental manipulation for the HCT is shown in Table 1.

4.2. ANALYSIS. Of the 50 participants recruited, 6 were excluded for low accuracy (accuracy rate less than 90%) on the attention checking control items, leaving 44 participants in the final analysis. Since each critical item appeared twice in the experiment (following the same design as Experiment 1 and 2), we constructed a Bayesian logistic generalized linear model to check if the scalar implicature rate differed depending on whether the participants encountered the item for the first or the second time and calculated the composite reliability of the critical items in the HCT. Then we combined the data of Experiment 3 with the data of Experiment 2, treating the TVJT, PST, and HCT as a between-subject manipulation. We constructed a Bayesian logistic regression model to explore how task variation influences scalar implicature computation. The probability of computing scalar implicatures was modeled as a function of task type (TVJT vs. PST vs. HCT), scale ("some-all" vs. "or-and" vs. "ad hoc") and their interactions. The model also included by-subject random intercepts and scale-by-subject slopes. The predictors were dummy coded with the "or-and" trials of the PST as the baseline for comparison. We also plotted participants' responses in the literal non-implicature trials across task and scale types.

4.3. RESULTS. For the experimental items, we again observed an interaction between lexical scales and task type that affected the estimated scalar implicature rate (Figure 4). When the "or-and" condition in the PST was set as the baseline for comparison, there was neither a statistical difference between the baseline and the "or-and" condition in the HCT (beta = 0.05, CI = [-6.12, 6.37]), nor between the baseline and the same scalar items in the TVJT (beta = 2.5, CI = [-4.17, 9.69]). For the PST, both the "ad-hoc" (beta = 18.01, CI = [9.8, 30.16]) and "some-all" items (beta = 32.98, CI = [19.01, 56.33]) elicited more scalar implicature computation than the baseline "or-and" trials, however, the estimated scalar implicature rate dropped significantly for the "ad-hoc" (beta = -22.47, CI = [-41.7, -10.31]) and "some-all" items (beta = -25.97, CI = [-48.8, -11.91]) in the HCT and in the TVJT as well. Same as the previous experiments, we did not find an effect of trial iteration on the estimated scalar implicature rate (beta = -0.24, CI = [-3.66, 3.14]). The coefficient alpha for the 18 HCT critical items was 0.91.

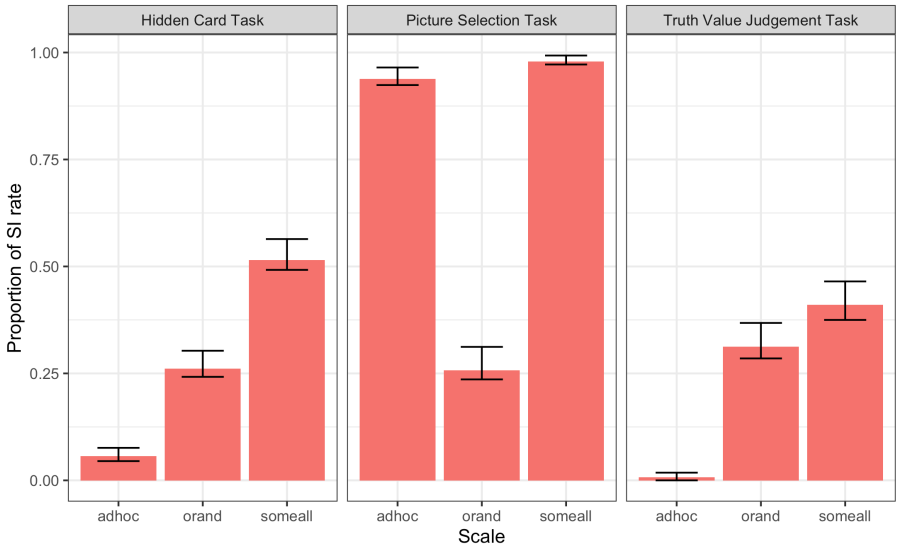

Figure 4: Rate of SI computation estimated by HCT, PST and TVJT in Experiment 2 and 3. The y-axis shows the percentage of deriving SI for a given scale ("ad hoc" vs "or-and" vs "some-all") in each task (HCT vs PST vs TVJT), with zero meaning zero percent and one meaning 100 percent. Confidence intervals were computed using bootstrapping methods.

5. Discussion. Though the TVJT, PST and HCT are supposed to measure the rate at which scalar implicatures are computed, the estimates from different tasks varied noticeably even for the trials that were parallel across the three tasks. The TVJT and PST were compared in a within-subject manipulation and the findings were later replicated in a between-subject manipulation. In general, the estimated SI rate was much higher in the PST than in the TVJT. This was the case for both the "ad hoc" and the "some-all" scale. When reading expressions like "Some of the animals are cats" or "The card has a cat", participants clearly derived scalar implicatures as shown in their strong preference for pragmatically felicitous cards in the PST, e.g. a card with three cats and three elephants or a card with only a cat. However, they still judged the pragmatic infelicitous interpretations (e.g. a card with six cats for the description "Some of the animals are cats") as "True" for the majority of the TVJT trials regardless of the scale. This pattern supported Katsos and Bishop's claim that participants are more tolerant to pragmatic infelicity than logical violations, and thus reluctant to judge under-informative statements as "False".

The form of the HCT appeared to mimic the PST in that there were two pictures provided and the only difference was that a visible card in the Picture Selection Task was replaced by the "Better Picture" card in the HCT; however, the estimated SI rate in the HCT was noticeably lower than in the PST for the "ad hoc" and the "some-all" scale. For example, while participants in the PST were more likely to select the card with only a cat rather than the card with both a cat and an elephant given the prompt "The card has a cat", they nevertheless preferred the card with a cat and an elephant when this card was paired with a "Better Picture" card in the HCT. The patterns observed in the HCT thus mimic the patterns in the TVJT more than those in the PST.

While "implicature rate" as a measurement construct was reliable in repeated measurements of participants within each task, it showed high variability and therefore, low reliability as a construct between the three tasks that we investigated in this study. These results suggest that either different tasks are measuring different constructs and "implicature rate" is not measuring the same thing across tasks, or that the pragmatics of each task is affecting the "implicature rate" in systematic ways that results in considerable variability across tasks. Most importantly, such variability in "implicature rate" between tasks suggests that comparing "implicature rate" between populations (e.g. children vs. adults) should also be approached with caution, since even within the same task, "implicature rate" may measure different things between populations or different populations may approach the pragmatics of each task differently. Future studies should also consider task variation and reliability across different populations and discover the sources of variability in experimental measurements of scalar inferences.

References

Barr, Dale J., Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language.* 68(3): 255-278.

Bott, Lewis and Emmanuel Chemla. 2016. Shared and distinct mechanisms in deriving linguistic enrichment. *Journal of Memory and Language*, 91: 117–140.

Bott, Lewis and Ira Noveck. 2004. Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language*, 51(3): 437–457.

Bürkner, Paul-Christian. 2017. brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1): 1–28.

Chemla, Emmanuel and Benjamin Spector. 2011. Experimental evidence for embedded scalar implicatures. *Journal of Semantics*, 28(3): 359–400.

Chierchia, Gennaro. 2013. *Logic in grammar: Polarity, free choice, and intervention*. Oxford: Oxford University Press.

Chierchia, Gennaro, Danny Fox, and Benjamin Spector. 2012. Scalar implicature as a grammatical phenomenon. In P. Portner, C. Maienborn, and K. von Heusinger (eds.), *An international handbook of natural language meaning*, Vol. 3, 2297–2332. Berlin: Mouton de Gruyter.

Crain, Stephen and Rosalind Thornton. 2000. *Investigations in universal grammar: A guide to experiments on the acquisition of syntax and semantics*. Cambridge: MIT Press.

Cronbach, Lee J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16: 297–334. http://dx.doi.org/10.1007/BF02310555.

De Neys, Wim and Walter Schaeken. 2007. When people are more logical under cognitive load. *Experimental Psychology*, 54(2): 128–133. https://doi.org/10.1027/1618-3169.54.2.128.

Degen, Judith and Michael K. Tanenhaus. 2015. Processing scalar implicature: A constraint-based approach. *Cognitive Science*, 39(4): 667–710. https://doi.org/10.1111/cogs.12171.

Degen, Judith and Michael K. Tanenhaus. 2019. Constraint-based pragmatic processing. In C. Cummins and N. Katsos (eds.), *The Oxford handbook of experimental semantics and pragmatics*, 485–504. Oxford: Oxford University Press.

Doran, Ryan, Gregory Ward, Meredith Larson, Yaron McNabb, and Rachel E. Baker. 2012. A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language*, 88(1): 124–154.

Feng, Shuo and Jacee Cho. 2019. Asymmetries between direct and indirect scalar implicatures in second language acquisition. *Frontiers in Psychology*, 10: 877.

Gamut, L. T. F. 1991. *Logic, language, and meaning. Volume I: Introduction to logic*. Chicago: University of Chicago Press.

Gazdar, Gerald. 1979. *Pragmatics: Implicature, presupposition and logical form*. New York: Academic Press.

Gerken, Louann and Mary E. Shady. 1998. The picture selection task. In D. McDaniel, C. McKee, and H. S. Cairns (eds.), *Methods for assessing children's syntax*, 157–173. Cambridge: MIT Press.

Geurts, Bart and Nausicaa Pouscoulous. 2008. No scalar inferences under embedding. In P. Egre and G. Magri (eds.), *Presuppositions and implicatures*, 1–22. MIT Working Papers in Linguistics.

Geurts, Bart and Nausicaa Pouscoulous. 2009. Embedded implicatures? *Semantics and Pragmatics*, 2: 1–34.

Gordon, Peter. 1998. The truth-value judgment task. In D. McDaniel, C. McKee, and H. Smith (eds.), *Methods for assessing children's syntax*, 211–228. Cambridge: MIT Press.

Grice, Herbert Paul. 1975. Logic and conversation. In P. Cole and J. L. Morgan (eds.), *Syntax and semantics 3: Speech acts*, 41–58. New York: Academic Press.

Grice, Herbert Paul. 1978. Further notes on logic and conversation. In P. Cole (ed.), *Pragmatics*, 113–128. Leiden: Brill.

Grodner, Daniel J., Nadine M. Klein, Kathleen M. Carbary, and Michael K. Tanenhaus. 2010. "Some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition*, 116(1): 42–55.

Hirschberg, Julia. 1985. A theory of scalar implicature. *Ph.D. dissertation*, University of Pennsylvania, Philadelphia, PA. http://repository.upenn.edu/dissertations/AAI8603648.

Horn, Laurence. 1972. On the semantic properties of logical operators in English. *Ph.D. dissertation*, University of California, Los Angeles, CA.

Horowitz, Anna C., Rebecca M. Schneider, and Michael C. Frank. 2018. The trouble with quantifiers: Exploring children's deficits in scalar implicature. *Child Development*, 89(6): e572–e593.

Huang, Yi Ting and Jesse Snedeker. 2009. Online interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. *Cognitive Psychology*, 58(3): 376–415. https://doi.org/10.1016/j.cogpsych.2008.09.001.

Huang, Yi Ting and Jesse Snedeker. 2011. Logic and conversation revisited: Evidence for a division between semantic and pragmatic content in real-time language comprehension. *Language and Cognitive Processes*, 26(8): 1161–1172.

Katsos, Napoleon and Dorothy V. M. Bishop. 2011. Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition*, 120: 67–81. https://doi.org/10.1016/j.cognition.2011.02.015.

Kuder, G. Frederic and Marion W. Richardson. 1937. The theory of the estimation of test reliability. *Psychometrika*, 2: 151–160. https://doi.org/10.1007/BF02288391.

Levinson, Stephen C. 1983. *Pragmatics*. Cambridge: Cambridge University Press.

Levinson, Stephen C. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. MIT press.

Noveck, Ira A. and Alicia Posada. 2003. Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language*, 85: 203–210.

Papafragou, Anna and Julien Musolino. 2003. Scalar implicature: Experiments at the semantics-pragmatics interface. *Cognition*, 86: 253–282. https://doi.org/10.1016/S0010-0277(02)00179-8.

Potts, Christopher, Daniel Lassiter, Roger Levy, and Michael C. Frank. 2015. Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. *Journal of Semantics*, 33(4): 755–802. https://doi.org/10.1093/jos/ffv012.

Rees, Anna and Lewis Bott. 2018. The role of alternative salience in the derivation of scalar implicatures. *Cognition*, 176: 1–14. https://doi.org/10.1016/j.cognition.2018.02.024.

Rees, Anna, Emily Carter, and Lewis Bott. 2023. Priming scalar and ad hoc enrichment in children. *Cognition*, 239: 105572. https://doi.org/10.1016/j.cognition.2023.105572.

Slabakova, Roumyana. 2010. Scalar implicatures in second language acquisition. *Lingua*, 120(10): 2444–2462.

Sperber, Dan and Deirdre Wilson. 1995. *Relevance: Communication and cognition*. Oxford: Blackwell.

Van Tiel, Bob, Eva Van Miltenburg, Natalia Zevakhina, and Bart Geurts. 2016. Scalar diversity. *Journal of Semantics*, 33(1): 137–175.