ASSESSING THE PRAGMATICS OF EXPERIMENTS WITH CROWDSOURCING: THE CASE OF SCALAR IMPLICATURES

There is a growing impetus to examine pragmatic phenomena experimentally (e.g., Breheney et al. 2006; Harris & Potts 2010). Potentially complicating these investigations is the way in which the experimental environment shapes participants' models of extra-linguistic context, or supplies 'demand characteristics' (Orne, 1962). Recent results indicate that the computation of pragmatic inferences, e.g., scalar implicatures, depends upon task structure (Huang et al. 2008), social norms (Bonnefon et al. 2009), and type of response elicited (Clifton & Dube 2010; Chemla & Spector t.a.). And yet these studies provide only a few points in a vast space of potential task parameters, thereby limiting our ability to systematically model the interaction between linguistic forms, context and pragmatic inference. Many incremental, parametric changes would need to be made to task instructions and situational context to uniformly sample the space. These manipulations, which are expensive and time consuming in the laboratory, become more tractable on crowdsourcing platforms.

In on-going work with Mechanical Turk we are exploring the task parameters that affect the rate of scalar implicature calculation in image verification. Our first study varied the presence and nature of preceding discourse context. We find that implicature rates in null contexts -- which almost all existing experiments employ -- match rates in contexts that are implicature defeating. Our second study changed task instructions to convince participants that they were performing quality control on prior Turker annotations, a manipulation which systematically lowered implicature-sensitive responses. This suggests that more strengthened interpretations may obtain in situations promoting evaluation apprehension (Rosenberg, 1965).

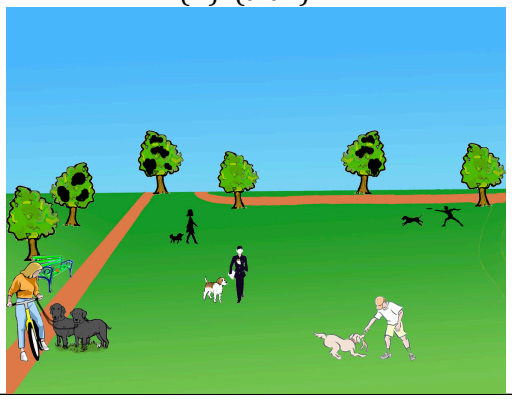
IMPLICATURE RATES FOR *SOME*

Standard error: 10-12% across cells. Significant effects: Instructions *($p < .01$),* Context *($p < 0.05$); by logistic mixed-effects modeling. No effects for* 'all' *descriptions (85% ± 8% accuracy) or fillers (93 ± 2%).*

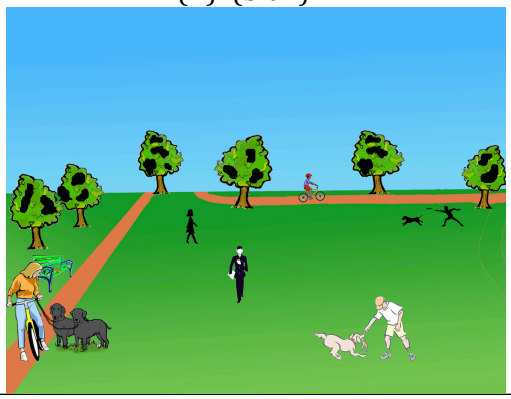
	CONTEXT TYPE			
INSTRUCTIONS	None	Implicature Defeating	Implicature Promoting	
Standard experiment	40%	40%	56%	45%
Quality control	23%	27%	38%	29%
	32%	*33%*	*47%*	37%

SAMPLE ITEM SET

(A): (a-all) T



(B): (b-all) T



(a) {Some, All} of the people have brought their dogs to the park.
(b) {Some, All} of the trees appear to have Black Spot Disease.

Condensed contexts for (a)

Implicature promoting: You are thinking about going to the park with your niece, but she is a little afraid of dogs. If too many people have dogs, you won't be able to go.

Implicature defeating: You are thinking about going to the park with your niece, but she is deathly afraid of dogs. If anyone has a dog, you won't be able to go.

REFERENCES

Breheney et al. (2006) Are generalised scalar implicatures generated by default? *Cognition.* **Harris & Potts (2010)** Perspective-shifting with appositives and expressives. *L&P.* **Orne (1962)** On the social psychology of the psychological experiment. *American Psychology.* **Bonnefon et al. (2009)** When some is actually all: Scalar inferences in face-threatening contexts. *Cognition*. **Clifton & Dube (2010)** Embedded implicatures observed. S&P. Huang et al. 2009. **Chemla & Spector (to appear)** Experimental evidence for embedded scalar implicatures. *JoS*. **Rosenberg (1965)** When dissonance fails: on eliminating evaluation apprehension from attitude measurement. *J Personality Social Psych.*