Introduction

1. Experiments are underspecified social interactions
   - subjects fill in missing social context from experience-based expectations [5]
2. Central Concern: How does experimental context influence participant responses?
3. Investigative Approach
   - Task focus: Image verification
   - Empirical focus: Scalar implicature rate for some (note wide rate variation in previous work: 34% to 93% [e.g., 1, 2])
4. 3 factors investigated
   - Linguistic context: specify the relevant bound at issue
   - upper bound [not all]
   - lower bound [at least some]
   - null context [target in isolation]
   - Evaluation apprehension [4]: anxiety of being judged
   - experimental protocol
   - annotation protocol
   - Task response categories [3]: structure of permitted responses
   - binary accuracy + Don’t Know
   - ternary informativity + False
5. Methodological conclusions
   - Implicature rate is affected by protocol type and response category
   - Response category interacts with linguistic context
   - but only rates for upper bound contexts are affected
   - Population effects may be large: university students and crowd-sourced workers differ
   - students sensitive to protocol
   - students and workers differently affected by linguistic context

Experimental Design

Task Instructions

Vary the social situation subjects believe they are participating in.

Experimental Protocol: Typical experiment instructions

Annotation Protocol: Checking the work of unaffiliated, possibly error-prone annotators

Context Story

All-Relevant - Upper bound is relevant; implicature is supported
All - Irrelevant - Upper bound is not relevant; literal interpretation is supported
None

Image Verification Task

Question: Would you be accurate if you said...

... ‘some of the spices have red lids’?

Available responses:
Yes, No, Don’t Know

EXPERIMENT 2 TASK RESPONSE CATEGORIES

Do specific prompt types affect implicature rates?

Yes - Response categories affect not only the baseline implicature rate but also interact with linguistic context.

Experimental protocol

Reliable effects:
context (p < .001)
prompt (p < .001)
context:prompt (p < .06)

Effects required for profanity is a mixed effects, logistic regression model (aiming to have a general effect on our data, e.g., overall implicature sensitive). The ANOVA results to show no effect of protocol

Annotation protocol

Not reliably different from experimental protocol

Post-experiment debriefing

subjects asked to comment on target trial responses

a majority show awareness of all alternative:
many indicated a preference for all alternative:

“ALL... would be a better response, but SOME... is true.”
“...all have nuts, not just some. I didn’t feel the categories fit.”

Acknowledgments

Thanks to Jay Padgett, and the LREC Semantics Lab audience for feedback. Many thanks to Ross Wilson for initial construction.

References