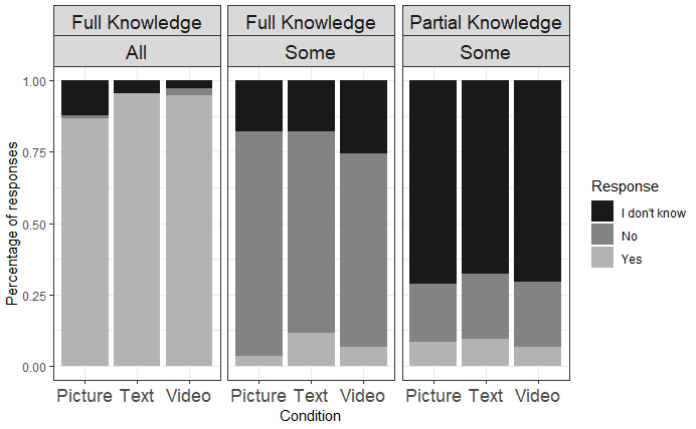
Sensitivity to speaker knowledge in online tests of scalar implicature

How is language comprehension impacted by how we experience contextual information? In two experiments, we asked whether online methods differed from in-person assessments of scalar implicature that relied on mental state reasoning - a task we reasoned might be especially sensitive to testing modality. We tested participants in one of four conditions: (1) in-person with a live-experimenter, (2) online with video stimuli, (3) online with pictures and text, or (4) online with text only stimuli. Across the experiments, no consistent differences emerged between modalities, suggesting that online methods provide valid measures of implicature under a variety of circumstances, even when relatively sophisticated mental state reasoning is involved. In particular, written stimuli were just as valid as video stimuli, if not more so.

Background: Research in semantics and pragmatics has recently witnessed rapid growth in the use of experimental methods that test large groups of participants to support robust statistical inference.[1-3] To facilitate this, many researchers have turned to online testing platforms such as Mechanical Turk and Prolific,[4-9] which include large groups of participants who speak diverse languages. However, while the validity of these online methods has been investigated in certain restricted test cases (e.g., acceptability judgments)[7-9] little is known about the impacts of online methods when testing pragmatic inferences, which often rely on subtle contextual parameters such as the knowledge states of particular speakers. For example, the computation of scalar implicatures (e.g., *some* implies *some but not all*) requires the hearer to assume that the speaker is knowledgeable about potential scalar alternatives. If contextual cues indicate that the speaker is not knowledgeable, hearers will derive an ignorance implicature instead (e.g., *some* implies *some and perhaps all*). What's unclear is whether such inferences about speaker states differ when an actual speaker, with actual mental states, is physically present vs. when a speaker is merely described via text, or otherwise represented via images or video. While implicatures have been documented across a variety of modalities,[1] it's unclear to what extent differences across these studies might be attributable to experimental modality. To investigate this question and probe the validity of remote testing methods, we tested participants' sensitivity to speaker knowledge when computing implicatures by presenting them with speakers across four modalities: in-person, remote video, remote photos, and text-only remote testing.

The Experiments: In Exp. 1, 90 English-speaking participants were recruited via Prolific, 30 per condition. These data were compared to existing data from 30 participants tested in-person by an experimenter who presented videos of the speaker on a laptop computer. Participants saw/read vignettes about a speaker (Mary) who had three boxes in front of her. Conditions differed in the modality: vignettes were presented either as video clips, still images, or short paragraphs of text. In each trial, the contents of the first two boxes were revealed to the participant and both always contained the same object types (e.g., apples). Mary then either looked inside the third box without revealing the contents to the participant, or did not look inside, and made a statement about the contents of the boxes using either 'some' or 'all'. There were therefore three types of trials: those where Mary looked in all three boxes and said 'all' (e.g., "All of the boxes have apples."; full knowledge/all), those where Mary looked in all three boxes and said 'some' (e.g., "Some of the boxes have apples."; full knowledge/some), and those where Mary looked in two out of three boxes and said 'some' (partial knowledge/some). Participants then answered a question about the contents of the third box (e.g., 'Do you think that there are apples inside the third box?'), by choosing "Yes", "No", or "I don't know".

Expected responses for each condition were as follows: full knowledge/all should lead to “Yes”, full knowledge/some should lead to “No” (as a result of computing a scalar implicature), and partial knowledge/some should lead to “I don’t know”. Participants completed a total of 9 trials (3 of each type). In Exp. 2, we conducted an exact replication of Exp. 1, but doubled the number of participants to 60 per condition, 180 total. The goal of Exp. 1 was to verify the reliability of effects observed in Exp. 1.



Results: Data from Exp. 1 were analyzed with the existing in-person data. We constructed a generalized linear model (GLM) predicting the proportion of participants’ “No” response to the trials with ‘some’ based on modality, knowledge state, and their interaction. The in-person condition was treated as the baseline. The model revealed a significant main effect of knowledge state (β=-2.84, SE=0.42, $p < 0.001$), as well as an interaction effect between modality and knowledge state; in particular, the proportion of “No” responses in partial knowledge trials increased with the online/video modality (β=1.25, SE=0.53, p=0.02; see Figure). As the expected response on partial knowledge trials was “I don’t know,” this effect suggests that participants in the online video condition were slightly more likely to compute scalar implicatures even though the speaker’s knowledge state (i.e., not knowing what is inside the third box) did not support doing so. In order to test whether participants were simply less attentive in an online setting, we reran the GLM model predicting the proportion of “I don’t know” responses. Shifting modalities from in-person to online did not result in an increase in these responses, suggesting that online participants were not overall less certain than in-person participants. For Exp. 2 we again created a GLM predicting the proportion of “No” responses to ‘some’ based on modality (picture vs. text vs. video). Contrary to Exp. 1, a chi-square test found no significant effect of modality (Deviance=1.49, df=2, p=0.47) with a larger sample size (n=60 per modality). A model predicting “I don’t know” responses found no significant effect of modality, replicating Exp. 1 (Deviance=2.12, df=2, p=0.35), again suggesting that modality did not affect participants’ attentiveness. **Conclusion**: We find no reliable impact of testing modality on how participants compute scalar implicature. Online text-only stimuli were just as likely to generate implicatures as richer modalities that featured images and video, despite the role of mental state reasoning in the tasks. **Refs**: **[1]** Chemla & Singh (2014). Remarks on the experimental turn in the study of scalar implicature, Pt I. L&LC. **[2]** Cummins & Katsos (2019). The Oxford Handbook of Experimental Sem. & Prag. **[3]** Devitt, M. (2011). Experimental semantics. P&PR. **[4]** Erlewine & Kotek (2016). A streamlined approach to online linguistic surveys. NLLT. **[5]** Munro et al. (2010). Crowdsourcing and language studies. **[6]** Fort et al. (2011). Amazon mechanical turk: Gold mine or coal mine? CL. **[7]** Sprouse, (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. BRM. **[8]** Schnoebelen & Kuperman (2010). Using Amazon mechanical turk for linguistic research. Psihologija. **[9]** Gibson et al. (2011). Using Mechanical Turk to obtain and analyze English acceptability judgments. L&LC.