A new Director task: Modelling common ground through referential specificity

Paula Rubio-Fernández (Massachusetts Institute of Technology; *prubio@mit.edu*) &
Julian Jara-Ettinger (Yale University; *julian.jara-ettinger@yale.edu*)

Over two decades, the Director task (DT) has increasingly been employed as a test of the use of Theory of Mind in communication, first in psycholinguistics and more recently in social cognition research. In this task, a participant follows the instructions of a confederate Director to move around various objects in a vertical grid of squares. The confederate sitting on the other side of the grid is ignorant of the contents of some of the cells because they are occluded on her side. A long series of studies has revealed that participants suffer ***interference*** from their privileged perspective when interpreting the Director's instructions (e.g., Keysar et al., 2000, 2003; Barr, 2008; Lin et al., 2010).

Keysar and colleagues and more recently social cognition researchers (e.g., Apperly et al., 2010; Dumontheil et al., 2010a, 2010b) have interpreted the poor performance normally observed in the DT as evidence of ***'limited use of Theory of Mind in communication'***. In this paper we challenge that conclusion and argue instead that the design of the DT itself is what limits participants' perspective-taking abilities, by imposing artificial demands on their selective attention. While seemingly uncomplicated, the design of the DT forces participants to supress a universal assumption in human communication: namely, that people know more than what they can see and are therefore able to refer to entities outside their visual field – unlike the Director.

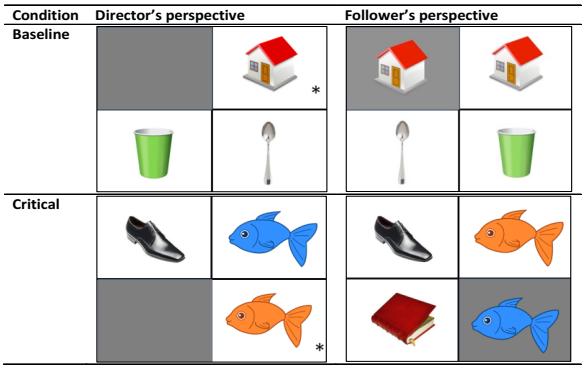
However, exclusively focusing on the objects that the Director can see in the grid need not be evidence of Theory of Mind use. In fact, the standard metrics of interference used in the DT cannot determine the extent to which participants' performance is dependent on Theory of Mind or selective attention. In other words, participants may suffer interference from their own perspective because they are failing to adopt the Director's perspective, or because they do not have enough executive control to inhibit their own. Thus, the aim of this study was to challenge the key assumption in the DT: that ***when participants consider the hidden objects as possible referents for the Director's instructions, they need not be failing to use their Theory of Mind***.

Pairs of naive participants played a new DT on two computers showing 2×2 grids of objects. One of the four cells had a grey background and contained an object in the Follower's grid, but was empty in the Director's (see Table 1). The Critical trials included a subtle manipulation: unbeknownst to the participants, the position of the grey cell was shifted in the Director's grid, so that a figure that appeared on a grey background in the Follower's grid now appeared on a white background in the Director's. The Director would see two fish of different colours, for example, and ask the Follower for 'the orange fish'. However, in the Follower's grid the blue fish appeared on a grey background, thus inviting the question: ***if the Director cannot see the blue fish, why does she call the target 'the orange fish', and not just 'the fish'?***

A direct measure of suspicion was collected in a post-test questionnaire and participants' fixations on the grey cell were taken as an indirect measure of suspicion. Only 4 participants (19%) reported not to have noticed anything peculiar in the Director's instructions. As predicted, those participants also said they had tried to focus their attention on the three white cells and block the grey cell from their view. According to the standard metrics of interference, those 4 participants would be 'model perspective takers'; however, those participants were actually underusing their Theory of Mind. By contrast, 16 participants (76%, sig. above chance) were suspicious that the Director sometimes knew about the contents of the grey cell because she used colour to distinguish the target. Also as predicted, those participants paid increasing attention to the grey cell in the second half of the task, when critical trials were administered and they grew suspicious of the Director's perspective. LMM comparisons confirmed that the suspicious participants' fixations on the grey cell

had a U-shaped distribution, first decreasing because of practice and then increasing because of suspicion (see Figure 1).

The results of this study confirm that ***the DT is not a reliable test of Theory of Mind use in communication***. According to the standard metrics of interference, 'optimal performance' is possible by using selective attention alone, while in this study the most sophisticated Theory of Mind use was revealed by those who kept track of the hidden cell.



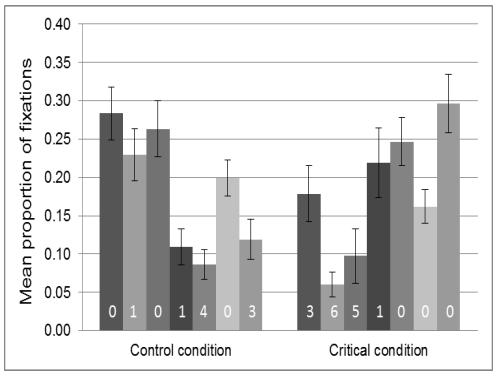


Table 1 (left): Sample trials from the Baseline condition and the Critical condition. The asterisk indicates the target object. The position of the objects was scrambled in the Follower's grid to avoid that the Director would use coordinates. **Figure 1 (right)**: Mean proportions of fixations on the grey cell in the seven trials of the Baseline condition and the seven trials of the Critical condition (by order of presentation from left to right) by those Followers who were suspicious of the Director's perspective (N=16). The number at the base of each bar indicates how many Followers did *not* fixate on the grey cell in that trial.

Following up on this study, we have tried to ***model common ground through referential specificity***. We presented MTurkers (N=40) with 4-figure displays, similar to those in the new DT. Participants had to estimate the likelihood that the speaker knew about the object in the grey cell given her instructions. To manipulate ***referential specificity*** we used **colour adjectives** ('the orange fish' vs. 'the fish'), **size adjectives** ('the small suitcase' vs. 'the suitcase') and **category level** ('the Porsche' vs 'the car'). To manipulate ***speaker's knowledge*** we used a **direct-reference condition** ('Select the blue fish'), a **hidden contrast condition** ('Select the orange fish', as in the new DT), a **hidden and shared contrasts condition** (adding a red fish on a white cell to the previous condition) and an **ambiguous condition** ('Select the fish' when there was a hidden blue fish).

We modelled the task through a speaker model that produces utterances (*Utt*) given a referent (*Ref*) by doing Bayesian inference over the listener's probability of recovering the referent given the utterance:

$$p_s(Utt|Ref) \propto p_l(Ref|Utt)p(Utt)$$

Our listener model jointly infers the referent (*Ref*) and the speaker's knowledge (*Cg*) by reasoning about the probability that the speaker would produce the heard utterance given different hypotheses about the speaker's knowledge, and the referent:

$$p_l(Ref, Cg|Utt) \propto p_s(Utt|Ref, Cg)p(Ref)p(Cg)$$