Replication of Study 1 by Keysar, Lin, and Barr (2003, *Cognition*)

rxdh@stanford.edu

Introduction

The idea that humans can accurately and effortlessly reason about the mental states of other humans is a foundational assumption of modern cognitive science. Among other things, this ability, called theory of mind, allows us to infer the underlying intentions that motivate others' actions, and to use these inferences to predict future actions. In this study, Keysar et al. argued that while adults are *capable* of applying this ability, they do not always apply it reliably.

To demonstrate the limitations, Keysar et al. used a director-agent task where one player gave instructions about how to move objects around a grid and the other player attempted to follow these instructions. For example, the objects may include a medium-sized measuring cup in one location and a smaller measuring cup in another location. The director may say 'move the large cup to the right,' presumably meaning the medium-sized one. Some objects were occluded such that only the agent could see them, creating a potential need for the agent to apply theory of mind and reason about the director's beliefs. For instance, if a much larger measuring cup was in an occluded slot and agents were not considering the director's beliefs, they might interpret 'the large cup' to mean this occluded cup even though the director didn't know about it. Indeed, the authors found that 74% of participants attempted to move this hidden item at least once (out of four critical cases) in the experimental condition, compared to 0% in a control condition where there was no ambiguity over the referent. 46% reached for it at least twice. These are our target findings.

Methods

Power Analysis

While no statistics were reported in the original paper, the original result concerns a difference in binomial probabilities q, where a `success' is attempting to move the hidden item. We expect $q \approx 0$ under the null hypothesis that people optimally use theory of mind (as found in the baseline condition), while the authors estimated an alternative rate of $q = .74$ from a sample size of $n = 38$. Note that the true statistical model underlying this binomial process is somewhat obscured, since this alternative rate actually reflects the fact that 74% of participants got a `success' at least 25% of the time.

We use the binom package in R to conduct a post-hoc power analysis of our ability to distinguish the estimated alternative rate from the null. This analysis is based on Wald statistics. Since power is theoretically not defined if $q = 0$ under the null hypothesis (because it would be impossible to empirically observe *any* successes), we set the null proportion to $q = .1$. The power of detecting a true rate of $q = .74$ is 80% for a sample of $n = 4$, 90% for a sample of $n = 6$, and 95% for $n = 8$. Although these small sample sizes would be sufficient to detect some effect, we propose a matched sample size of $n= 38$ in order to get a reliable re-estimate of the effect size.

Planned Sample

We will recruit 38 participants on Amazon mechanical turk. We will restrict participation to the U.S. and will exclude participants who report confusion about the task instructions in an exit survey or make errors on more than 10% of the non-critical instructions (which serve as catch trials).

Materials

The original article states:

> "Participants sat at a table opposite a confederate director. On the table was a vertical array of 4 x 4 slots. Five of the slots were occluded from the director's perspective, and the remaining 11 were visible to both participant and director. One of these unoccluded slots included an object that was the target, such as a cassette tape box. Another object, such as a roll of tape, was hidden by the participant in a small brown paper bag and placed in an occluded slot. In addition to the intended object and the object in the bag, each array had several unrelated objects. For each grid, there was one "critical instruction" (e.g. "move the tape") in which the director gave an instruction to move a mutually-visible object that could also potentially refer to the object hidden in the bag.
>
> The experiment had eight items, each with a different set of objects and critical instruction. Each item consisted of a series of instructions to move objects around and included one critical instruction. Each item also included one critical pair of objects, one of them the intended object and the other hidden in the bag, such as the cassette tape and the roll of tape.
>
> To collect baseline performance information for each item we added a condition in which the hidden object (e.g. roll of tape) was replaced with an object that did not fit the critical instruction (e.g. a battery). Thus, the `move the tape' instruction appeared after the participant had hidden in the bag either a roll of tape (experimental condition) or a battery (baseline condition). Each participant received half the items in the experimental condition and half in the baseline condition. Items and conditions were counterbalanced across participants. Order of presentation was random, with the provision that no more than two items in the same condition would appear consecutively.
>
> A trained female confederate played the role of the director in order to ensure uniformity of critical instructions across conditions and participants. The confederate was well practiced in playing the role of a naive participant. To create a realistic situation, she indicated having some difficulty with the task, interjected her instructions with hesitations, and made occasional errors with non-critical objects. In addition, the director improvised most of the instructions, except that critical instructions for the target objects were scripted. Indeed, with the exception of one person, none of the participants later reported that they suspected during the experiment that the director was a confederate."

We will precisely replicate all objects sets, instruction sets, and randomization schemes. However, given the constraints of the web experiment format, there will be several exceptions as well. First, participants cannot be seated across from each other at a table: they will each see their respective views of the 4 x 4

grid on a screen and will communicate via a text box. Second, we will not use a trained confederate, and will instead randomly assign one of the players to the role of the instruction giver. To make sure the critical instructions are still scripted, we will provide a "send instructions" button for a subset of trials (including some non-critical ones).

Procedure

The original article states:

> "The participant and the confederate arrived at the laboratory and the experimenter explained that they would be playing two different roles in a communication game. She then assigned roles, ostensibly randomly, and the participant received the role of the `addressee' while the confederate was assigned the director's role. At the beginning of each item, the director received a picture of the array of objects with arrows indicating where each object should be moved. The arrows were numbered to specify the order of object movement. The director then used this picture and instructed the participant to rearrange the objects accordingly.
>
> The director's picture showed her perspective, meaning that only mutually-visible objects appeared on the picture, with the remainder of the slots clearly occluded. This was demonstrated to the participant, and the experimenter also pointed out that the objects in the occluded slots were not part of the game. In addition, before each item began, the experimenter put a large cardboard wall between the confederate and the participant as a visual barrier. Then she handed the participant an object and a brown paper bag and asked the participant to hide the object in the bag and place the bag in one of the occluded slots. The experimenter did not name the object but simply handed it to the participant and referred to it only as `this'. After the participant had hidden the object in the bag and put the bag in the slot, the experimenter removed the barrier and the director started with the instructions.
>
> The experiment began with a practice item to familiarize the participant with the task and to correct any misunderstanding. In order to make absolutely sure that the participant fully appreciated the director's difference in perspective, the participant and the director switched roles and the participant gave instructions for a second practice item. In this manner there would be no question that the participant understood the information provided in the picture of the array, appreciated that the director could not see hidden objects, and knew that the only objects relevant to the game were the mutually-visible ones. After the role reversal, the participant and the confederate resumed their original roles and the experimenter presented the first item. The experiment proceeded through all eight items, with the director providing instructions and the participant moving the objects. Before each instruction the director said `ready?' at which point the participant looked at the center of the array and answered `ready'. The participant was free to converse with the director, to ask questions and so on."

We will follow these procedures precisely, but all the practice items will be moved to a general set of instructions, before players enter the real-time, multi-player environment. Because there will be no

experimenter actively administrating the practice trials, we will give a comprehension test to participants before they are assigned to the role of 'director' or 'agent.' In this way, both participants will understand both tasks and appreciate the information asymmetry.

Analysis Plan

To analyze the percentage of participants who 'attempted to move the hidden item' at least once, we must define what it means to 'attempt to move' an item. The authors did not provide criteria in their original report. We will therefore propose two different analyses. The first (and roughest) will be based on error data: if a player clicks and drags some object other than the target, or drags the target to the incorrect destination, we will log the error, correct it, and give the participant another chance to get it correct. We can tally up the number of times a participant moves the hidden distractor instead of the target, and if this is positive, it will be considered consistent with the results of Keysar et al (2003).

Since this is a very conservative definition of 'attempting to move,' it is unlikely a priori that participants will register at all on this scale. Thus, we turn to mouse-tracking data which more accurately matches the measures used in the paper: simply reaching toward a hidden distractor was counted as an 'attempt.' We can sample the matcher's mouse movements every 15ms and can therefore compute standard mouse-tracking dependent measures. We plan to focus on the simple measures: (1) hover time -- how long, if at all, did the player's mouse hover over the distractor -- and (2) normalized proximity -- was there a time window in which the mouse was getting closer to the distractor yet farther away from the target?

Differences from Original Study

The primary difference between our study and the original is the fact that it will be run online with participants connected via a virtual environment, instead of face-to-face in a room. A related difference is our decision not to use a confederate. In general, studies using confederates are difficult to replicate exactly and are impractical to run online. We are aware that this change (while keeping scripted instructions, which were the primary justification for the confederate) may lead to additional variation across groups. It is also known that textual communication (via our instant messaging interface) can differ from verbal communication (as in the original study). Finally, the way people attempt to move things in our virtual environment may be different from how people attempt to move things in the lab, and we cannot eye-tracking data to back up our analysis, as the original paper did.

(Post Data Collection) Methods Addendum

Actual Sample

We recruited 34 participants (17 pairs) from Mechanical Turk. All participants were from the U.S., as planned. Three pairs were excluded for making 2 or more errors on non-critical items.

Differences from pre-data collection methods plan

Technically 2 errors on non-critical items is an 8% rate (2/(32-8) = 8.3%), but we considered this to be

close enough to our pre-planned 10% rate. We were not able to reject people based on confusion as planned, because we did not post an exit survey.

Because we were not able to obtain exact copies of materials from the original authors of the study, we had to create our own filler items and instruction sets (available here). We did, however, keep the original study's *critical items and distractors*, their scripted instructions for those items, and also their randomization schema.

Finally, we did not put hidden items in a paper bag, as in the original study.

Results

Data preparation

First, we excluded three participants based on the number of errors they made on non-critical (filler) items. An error was recorded if a participant clicked on the wrong object and dragged it anywhere, or if they clicked on the correct object and dragged it to the wrong location. Although these errors are always based on agent behavior, note that they could plausibly be the result of faulty director instructions.

Confirmatory analysis

Using the error rates on *critical* items, we reconstructed the target result (Table 1) from Keysar et al (2003). We find that 93% of participants attempted to move a hidden object at least once in the Experiment condition, out of four possible items, compared to only 14% in the baseline condition. This is similar to the effect observed by the authors in the original study, which found 71% and 0%. Our errors were larger across the board, perhaps due to the interface and the Turk population, but the gap between the two is about the same size.

	At least once	At least twice
Experiment	93	64
Baseline	14	0

Exploratory analyses

Item-level differences

First, there is some concern over the choice of the 8 critical items. We were interested in how the error rates were distributed across different items. Of all the errors participants made critical items in the experimental condition (i.e. the errors we counted in the 93% statistic reported above), we counted how many of these errors occured in each item:

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8

# incorrect	1	5	1	3	2	6	6	1
# correct	5	3	1	8	2	2	4	6

We see that over 60% of groups that were assigned to the experimental condition of items 2, 6, and 7 made errors, but less than 30% made errors on item 1, 4, and 8. A χ^2-test for independence did not reject the null hypothesis that error rate and item number are independent, $\chi^2(7) = 10.89, p = 0.14$, but we return to this point in the discussion.

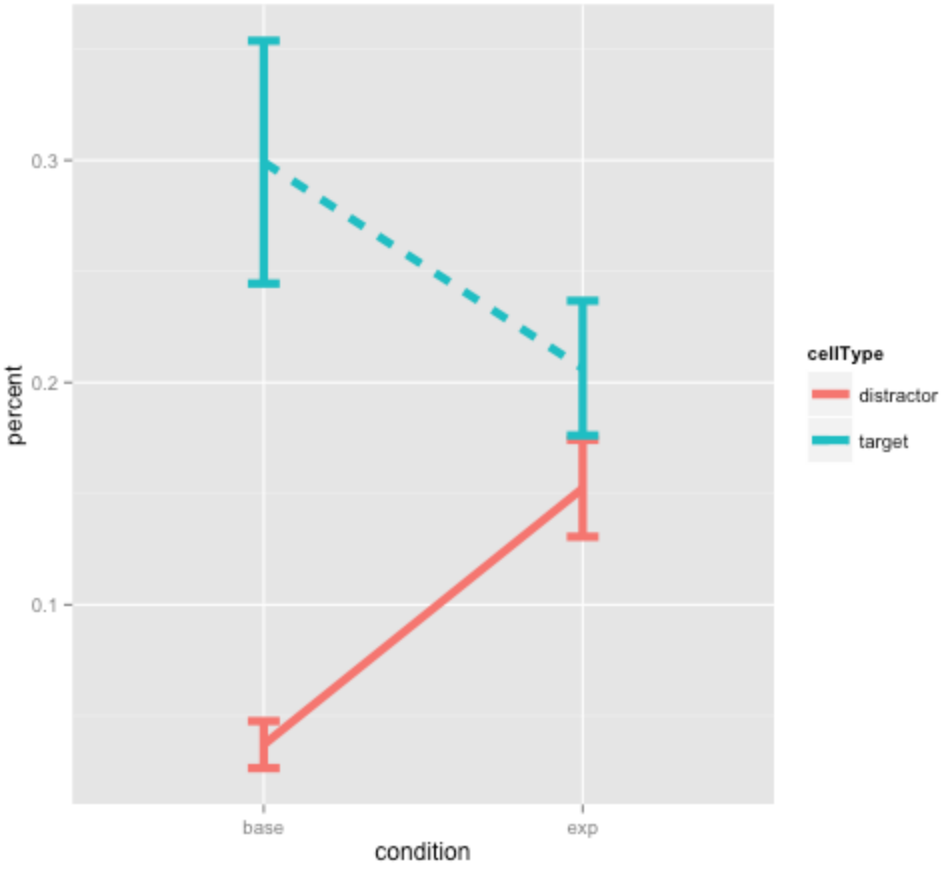
Mouse-tracking analysis

Second, we were interested in using a mouse-tracking analysis to follow-up on the raw error analysis reported above. This is meant to mimic the spirit, if not the precise content, of the eye-tracking analysis reported by Keysar et al (2003).

We define the decision window as the span of time between the point when the matcher received their instruction message and when they started moving an object. If it took them multiple attempts to move the correct object, we restricted our analysis to the first attempt. We took the set of all mouse position samples in the decision window (collected approximately every 15ms), and measured what proportion were located inside the cell containing the target vs. the cell containing the distractor.

We had to exclude an additional 3 participants for this analysis, because the timestamps for director and matcher did not align and we could not establish the decision window properly -- if we ran it again, we would index mouse movements the server time, but we accidentally used the local time stamps that players sent along with their messages.

We found that there is an interaction between cell type and condition on the percent of hover time spent in that cell, $F(3, 52) = 10.7, p < 0.001$. In baseline trials, people spent much more time in the target cell and much less time in the distractor cell than in the experimental condition, with a significant interaction coefficient, $b = -0.21, t(11) = -3.1, p = 0.003$. See the figure below. Of course, this also includes participants who actually made errors, since the data is too sparse to exclude them.



Discussion

Summary of Replication Attempt

In Keysar et al (2003), the authors found that 71% of participants attempted to move a hidden object at least once when given a potentially ambiguous instruction, compared to 0% in the baseline condition, therefore failing to take into account the director's perspective. We found 93% in the experimental condition and 14% in the baseline condition using raw error counts. We also found that people were more likely to hover over the distractor in the experimental condition than the baseline condition. Although there were a number of potentially concerning differences between our online version and the original version, it's fair to say we replicated the original results.

Commentary

First, our analysis of item-level differences shows that the dependent variable used in the original study (i.e. "percentage of participants who moved the critical item at least once") is somewhat problematic. It could look like 100% of participants made the errors even if they all made those errors on one particularly difficult item (e.g. the 'bottom' block, where they actually mean the one on the

second-to-bottom row). If we exclude the three 'hard' items (2, 6, and 7) where over 60% of participants in the experimental condition made errors, our results look much less strong than Keysar's.

	At least once	At least twice
Experiment	43	14
Baseline	7	0

This suggests that (1) the "at least one error" DV is not appropriate in settings with high variability across items and (2) we might want to be more careful about controlling for this variability, perhaps by first measuring salience and ambiguity *without* the occlusion aspect of the paradigm.

Second, it's worth thinking about some ways our online version may differ from the original. The fact that we didn't use a confederate didn't seem to be a problem, given our use of randomly scripted instructions. Similarly, because players could type messages directly to one another, and since the director could watch the matcher moving objects in real-time, it's also fair to say that participants truly believed they were playing with another human being.

One major concern is our graphical representation of occluded cells. We simply told players that items behind black cells were hidden to the director, and this was a critical question in the quiz that players had to pass before playing, so we know that they were aware of it. But they may forget part-way through the experiment, or don't fully internalize it. In the in-lab version, there was an actual divider between players, so it would have been more salient. This would mean that some errors in our version game aren't due to a failure of perspective-taking, but instead because of a misunderstanding of what the other person's perspective is to begin with.

Finally, based on the free production of the director in the absence of scripted instructions, I'd like to offer an alternative explanation for these results (which also applies to Kesyar). The director tends to be naturally *over-informative* with their freely produced instructions. Instead of "move the stuffed animal down", they'll say "move the stuffed panda bear down." Or, instead of saying "move the plane to the right," they'll say "move the red airplane to the right," even though there is only one plane. This is strategic, since the director *knows* there are hidden objects they don't know about.

If the matcher assumes that the director will behave in this optimal, over-informative manner, thereby successfully applying theory of mind, then they wouldn't expect such ambiguous instructions to be produced. Ironically, this could mean that *successful use of theory of mind* actually causes the apparent *failure* of theory of mind in this paradigm because of the artificially ambiguous instructions we force the director to use. It would be interesting to rerun these experiments without scripted instructions and test whether errors go down. Along the same lines, we could test what words people naturally use to refer to these objects outside the context of the experiment (i.e. their "basic-level" labels), and compare them to the words people use inside the experiment.