Replication of Experiment 1 by Kovács, Téglás and Endress (2010, *Science*)

Desmond C. Ong
Department of Psychology, Stanford University
desmond.ong@stanford.edu

Introduction

The original paper by Kovács, Téglás and Endress (2010) (Supplementary materials) showed that both adults and 7-month-old children automatically code others' beliefs in a modified "false-belief" task. This encoding affected participants' own beliefs, even when the other's beliefs were irrelevant and when the other is absent.

The target study for replication will be Experiment 1 from the paper, which had N=24 adults tested in the lab. Participants were shown a movie of an agent watching a ball roll around on a table, either staying behind an occluder or moving off the screen. In all the conditions, the agent leaves for a short duration, during which the ball either remains where it is or moves, thus introducing the "false-belief". The agent returns, and the occluder is removed; participants are asked to respond when they have identified the location of the ball. The reaction times were taken as the dependent variable in this experiment. While the agent's beliefs were irrelevant to the task, they still affected the reaction times, and the authors interpreted this as evidence of automatic encoding of others' beliefs.

The experiment involves a 2 (Participant believes ball is behind occluder "P+", or not, "P-") x 2 (Agent believes ball is behind occluder "A+", or not "A-") x 2 (Ball is actually behind occluder "B+", or not, "B-") design. Hence, there are eight conditions, and an example would be P+A+B+. When the ball is not present, however, there is no reaction time associated with detecting the ball (the correct answer is not to respond); reaction times in the original paper were only reported for the "B+" trials. Each participant watches each movie 5 times, for a total of 40 movies.

The main findings of Experiment 1 of the study are:

1) Reaction times to detect the ball after the occluder drops when both participants and agents believe that the ball is behind the occluder [P+A+] is significantly faster than the baseline, when neither participants nor agents believe that the ball is behind the occluder [P-A-]

2) Reaction times when only the participant believes the ball is behind the occluder [P+A-] is similarly significantly faster than the baseline [P-A-]

3) **The critical finding**: reaction times are faster when only the agent believes the ball is behind the occluder, but not the participant [P-A+], as compared to the baseline when neither believe the ball is behind the occluder [P-A-]. This, they interpret, as evidence for automatic encoding of others' beliefs.

4) Finally, they find no significant differences between reaction times when only the

participant believes the ball is behind the occluder [P+A-], and when only the agent believes the ball is behind the occluder [P-A+].

These findings are also summarized in Table 1, below.

Methods

Power Analysis

Assuming that the original effect size is **0.450** (see Table 1, below: the third comparison, P-A- with P-A+, was the so-called "critical" comparison), for a two-tailed matched-pairs comparison with alpha = 0.05, we would require:

For 80% power, N=41
For 90% power, N=54
For 95% power, N=67

I would plan to collect N=60 participants to achieve >90% power.

Comparison group 1	Comparison group 2	t-statistic	df	P value	Cohen's d	Effect Size
P-A-	P+A+	3.47	23	0.002	1.447	0.586
P-A-	P+A-	3.43	23	0.002	1.430	0.582
P-A-	P-A+	2.42	23	0.02	1.009	0.450
P-A+	P+A-	0.99	23	0.33 n.s	0.413	0.202

Table 1: Original comparisons (p. 1832). The t-statistic, dfs, and p values are given in the paper. Cohen's d and effect sizes are calculated from the given t-statistics and dfs.

Group	Mean (ms)	SEM (ms)	N	SD = SEM*sqrt(N)
P+A+	313	10	24	31.62
P-A-	360	14	24	44.27
P-A+	328	12	24	37.94
P+A-	320	10	24	31.62

Table 2: Group level data, estimated from Fig 1A from the original paper. Note that the t-statistics reported in the paper (reproduced in Table 1) are probably from a matched-pairs t-test, as the degrees of freedom is 23, and so it is impossible to reproduce those t-statistics from the group-level data presented here.

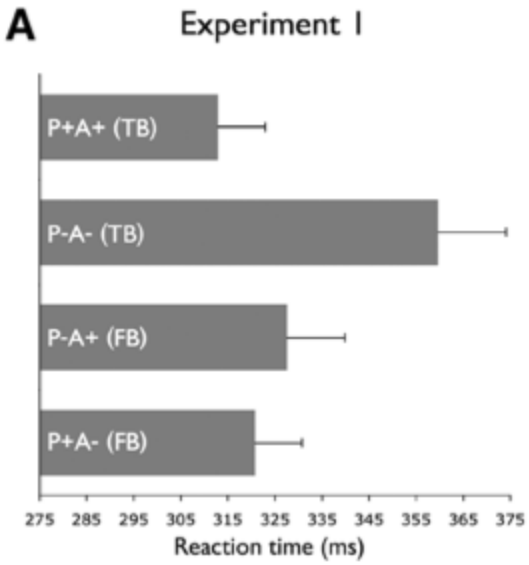

Figure 1: Figure 1A, reproduced from Kovács' et al (2010)

Planned Sample

Planned sample size: 60
Approximate task duration : 20-25 mins
Cost per participant: $1.50/participant ($3.6-$4.5 hourly rate)
Total cost: 60 * $1.50 = $90
Pre-selection rules: NA
Demographic restrictions: NA

Materials

"Participants watched 40 18.4s long animated movies, generated using Maya 3D software..." (From original article)

Stimuli movies were made by Andrew Surtees as part of an independent attempt to replicate the same paper. These movies are slightly longer than the original (25s), but followed the same patterns as the stimuli described in the paper.

Procedure

"In the last scene of the movie, the agent returned and the occluder was lowered; participants were instructed to press a button as soon as they detected the ball after the occluder was lowered. We recorded their reaction times. " (From original article)

The stimuli were embedded in a HTML webpage using HTML5's video tag. Javascript was used to randomize the stimuli presentations, and to record the reaction times using the

keyboard. As in the original experiment, participants were asked to press a button when the agent was detected leaving the scene (in our case, the letter “q”; attention check), and another button once the occluder dropped and the participant has detected the ball (in our case, the letter “p”; main dependent measure). Reaction times were captured for both detection tasks.

The original article was not clear on how “reaction times” would be defined for trials where the ball was not present. It seems like only the reaction times for the trials where the ball was present were analyzed.

The original article also did not define what counts as passing the attention check, and what is required to answer a “ball absent” trial correctly. In this replication, we have decided on a time limit of 3s: participants pass the attention check if they indicate within 3s of the agent leaving, and participants would answer a “ball absent” trial correctly if they wait for 3s after the occluder drops without pressing the response button.

Analysis Plan

The reaction times for detecting when the agent leaves the scene (attention check) and when the ball was detected (main dependent measure) are both measured.

- Participant exclusion rule: Exclude participants who:
    - drop more than 5 trials out of 40; or
    - report any technical difficulty.
- Data exclusion rule: Trials where the participant either
    - fails the attention check (takes longer than 3s to detect agent leaving); or
    - gets the trial wrong (i.e. presses button when ball is absent, or taking longer than 3s to press the button when ball is present).

After data cleaning:

1. For each participant, average reaction times across the repetitive trials (there are 5 repetitions per condition).
2. For each of the 4 comparisons listed in Table 1 above, perform a matched pairs (within-subject, across-conditions) t-test.

Differences from Original Study

1. Differences in sample size: original study had N=24, planned replication will have N=60
2. Original study was run in a lab; replication will be run online using Amazon Mechanical Turk.
3. No anticipated differences in results.

(Post Data Collection) Methods Addendum

Actual Sample

Actual sample size collected = 60, with final size 54 after exclusions.

Participant exclusion rule: Original criteria was too stringent (it would exclude 25% of the sample, and corresponded to about a 75/80 (including both ball trials and attention checks) or ~94% accuracy). Criteria was lowered to >90% accuracy. 6 subjects failed this criteria and were dropped. Final sample size was 54, which still achieves 90% power.

Data exclusions: based on analysis plan, individual trials which participants failed attention checks or were inaccurate at detecting the ball were dropped.

Differences from pre-data collection methods plan

Participant exclusion criteria had to be made slightly less strict. Note that the original study did not specify any participant inclusion criteria. Other than that, no differences.

Results

Data preparation

Data preparation was done as detailed in the analysis plan.

Confirmatory analysis

The same analyses as reported in the original paper were calculated for replication. The critical replications are presented in Table 3 (to be compared with data from the original, in Table 1). Additional information about the group level statistics are provided in Table 4 (to be compared with Table 2), and in Fig. 2 (to be compared with Fig. 1).

Overall, there is a systematic difference of about 300 milliseconds between all of the replicated reaction times and the reaction times reported in the original paper. This could be due to differences in deciding when to start the clock, i.e. we started the reaction time when the fence began to fall, while the original paper could have started the clock once the fence has fully fallen. This does not change the general trend of the results, though.

One aspect in which our data differs from the original is that the condition in which participants are supposed to be the fastest (P+A+) does not happen to be the fastest in the replication.

Group 1	Group 2	t-statistic	df	P value	**Replication**	Cohen's d	Effect Size
P-A-	P+A+	2.086	53	0.0418	**Yes**	0.573	0.275
P-A-	P+A-	4.487	53	<0.001	**Yes**	1.233	0.525
P-A-	P-A+	4.366	53	<0.001	**Yes**	1.199	0.514
P-A+	P+A-	0.579	53	0.565 n.s	**Yes**	0.159	0.079

Table 3: Replicated comparisons (to be compared with Table 1).

Group	Mean (ms)	SEM (ms)	N	SD = SEM*sqrt(N)
P+A+	740.08	28.03	54	205.94
P-A-	791.87	28.67	54	210.71
P-A+	684.72	18.89	54	138.84
P+A-	674.59	19.63	54	144.23

Table 4: Replicated group level data (to be compared with Table 2).

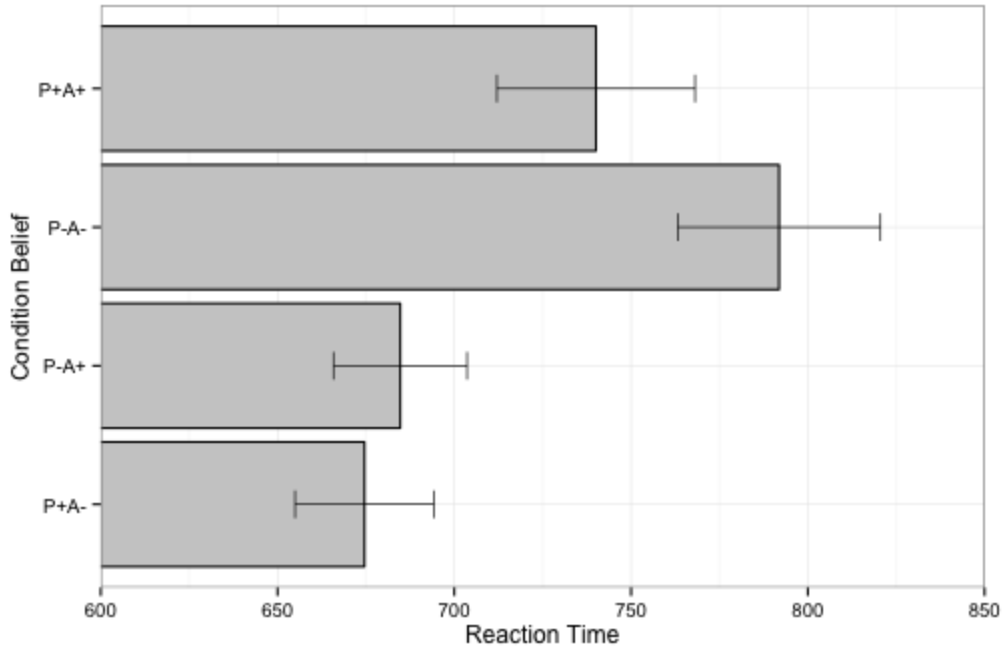

Figure 2: An analogous graph to the original, reproduced in Figure 1.

Discussion

Summary of Replication Attempt

The primary result from the original paper was a reduction of people's reaction time to detect an object when an (irrelevant) agent had a true belief about the object's location, even when they (the participants) did not. This implies that we automatically encode beliefs of others in situations, even when their beliefs are not necessary to the task at hand.

The results from Experiment 1 of the original paper were successfully replicated in this study, with an approximate power of 90%. In particular, the four statistical t-tests reported were all replicated.

Commentary

The findings seem to be robust enough to be replicated on an online sample, despite the obvious (and real) concern for distracted participants.