# 1 Summary of data analysis draft

## 1.1 Results

Our results show that participants generally go to the canteen at 8:40 and earlier (around 90% of the time), and generally the office at 9:00 and 9:10 (90% of the time), while around 40% chooses the canteen at 8:50. These results show that while participants do consider the possible arrival times for the other player, they only infer whether both players arrived early enough to go to the canteen, while they do not infer that the other person might have differing beliefs, causing them to choose office. In other words, participants do not seem to substantially consult their Theory of Mind in the Canteen Dilemma.

The data concerning their estimated certainty of success tells a different story. While participants go to the canteen 90% of the time at 8:30 and 87% of the time at 8:40, i.e. a 3% drop, when it comes to certainty, 70% of these participants are are very certain that the other player also chose canteen at 8:30, while only 51% are very certain at 8:40, i.e. a 19% drop. This indicates that participants who arrive at 8:40 do consider and reason about the other player's beliefs and intentions to a larger degree than is reflected in their choices. This is important for experiments concerning Theory of Mind, since it shows that there might be different agent types even among those making the same choices, in the sense that their latent Theory of Mind might make them certain to different degrees.

## 1.2 Reservations about results

### 1.2.1 Why do participants go to the canteen at 9:00 and 9:10?

Possible explanations include:

**(i)** participants chose at random.

**(ii)** Participants believed they could go to the canteen at 9:00 or earlier, or

Now consider the instructions given:

> "Both of you like to meet in the canteen for a coffee. If you arrive before 9:00 am, you have time to go to the canteen, but you should only go if your colleague goes to the canteen as well. If you or your colleague arrive at 9:00 am or after, you should go straight to your offices."

We can reasonably expect random or semi-random answers in any experiment, no matter the simplicity of the rules (i.e. people answering simple questions wrong does not necessarily imply a cognitive deficit, given the possibility of either clicking a wrong button unintentionally, or simply choosing at random). So (i) must be part of the explanation. The question is whether (ii) might hold. Participants might have only read the first sentence, interpreted 'before 9:00 am' as before or at 9:00 am, or interpreted 'you should go straight to your offices' as a moral imperative which might not align with their self-interest. Of those three, I take the last disjunct to be most likely, but it seems most plausible that (i) is the leading cause for the following reason.

### 1.2.2 Why do participants choose canteen at 8:50 (or earlier)

The rules state "If you arrive before 9:00 am, you have time to go to the canteen, but you should only go if your colleague goes to the canteen as well. If you or your colleague arrive at 9:00 am or after, you should go straight to your offices."

The rules intend to communicate: (a) you prefer to go to the canteen, but only with your colleague, (b) you have to go to the office if you or your colleague's arrival time is 9:00 or later, and (c) you always

arrive within 10 minutes of each other. It is also possible that the rules do not clearly enough state that it is far preferable to go to the office together, than to go to different places. The problematic interpretation by participants is **(d)** we both like to meet in the canteen, and we do not like if one or both goes to the office. This, together with the rules (b) that you have to go to the office when one player arrives 9:00 or later, would entail that if a player arrives at 8:50 and thinks the other player arrives at 8:40, the only winning play is canteen. However this is a peculiar interpreation, since (a) entails that participants should only go to canteen when the other players do so. Now let us consider possible explanations for going to the canteen at 8:50.

**(i)** Participants understood (a) but either not (b) or (c), i.e. they ignored the possibility that the other player would arrive at 9:00 (this is arguably a 0-order ToM mistake, i.e not inferring (b) or (c) from the rules, but up to discussion if it is a 1st-order ToM mistake)

**(ii)** Participants understood (a), (b) and (c) and that the other player might have arrived at 9:00, but simply took a risk and went to the canteen at 8:50. This is somewhat reflected in certainty estimates given by participants.

**(iii)** Participants did not understand either (a), (b) or (c), implying that even if they considered the other playing arriving at 9:00, they thought canteen would still be the right choice. Like those going to the canteen at 9:00, it might be due to only focusing on the 'Both of you like to meet in the canteen for a coffee', while also understanding 'If you arrive before 9:00 am, you have time to go to the canteen' and ignoring the rest.

**(iv)** Mistakes/random choices. If some choose canteen randomly or by mistake at 9:00, we would expect some to do the same at 8:50. But this does not account for the relative difference between arrival times.

Explanation (ii) implies that participants makes choices where they know there is a chance of failure, which is a plausible assumption. In fact, if a participant knows (a), (b) and (c), and assume that another player would always go to the canteen at 8:40, it entails that it would be rational to go to the canteen half of the times they arrive at 8:50. This would explain the canteen choices we see at 8:50, and also explain the dominance of canteen choices at earlier times. A participant might believe that other players go to the canteen at 8:40 if (1) they do not believe that other players consult their ToM, or (2), they might believe that while other players might have ToM considerations, they do not believe others do, so they go to the canteen at 8:40 none-the-less. In other words, if the importance of ToM is not common knowledge, it is not strictly rational to act on it.

We can also consider the possibility that those arriving at 8:50 did not believe that the other person would necessarily choose office at 9:00. While irrational, our data shows it's a true assumption. If a player have this belief, and the belief the other person would always choose canteen at 8:40, it would entail canteen as the rational choice at 8:50. This is an implausible explanations for canteen choices at 8:50. First, it ascribes implausibly complex reasoning to participants. Secondly, participants still go to the canteen at 8:50, even if we only include those who chose office at 9:00. In other words, they do not choose canteen at 8:50 because they themselves choose canteen at 9:00.

### 1.2.3 Why might participants choose office at 8:50, and earlier?

Participants choosing canteen 50% of the time they arrive at 8:50 means the same as choosing office 50% of the time, and the explanation for this is the same as above. The reason for preferring office at 8:50 over canteen would be due to consulting one's Theory of Mind. Participants who choose canteen at 8:50 will lose at least half of the times no matter what, while their other option, office, could win 100% of the time (ignore difference in payoff for now) if the other player chooses office as well. A participant arriving at 8:50 might think the following. Even if the other player arrives at 8:40, that player might not choose canteen, because that player knows that if the other player arrived at 8:50, they might choose office..

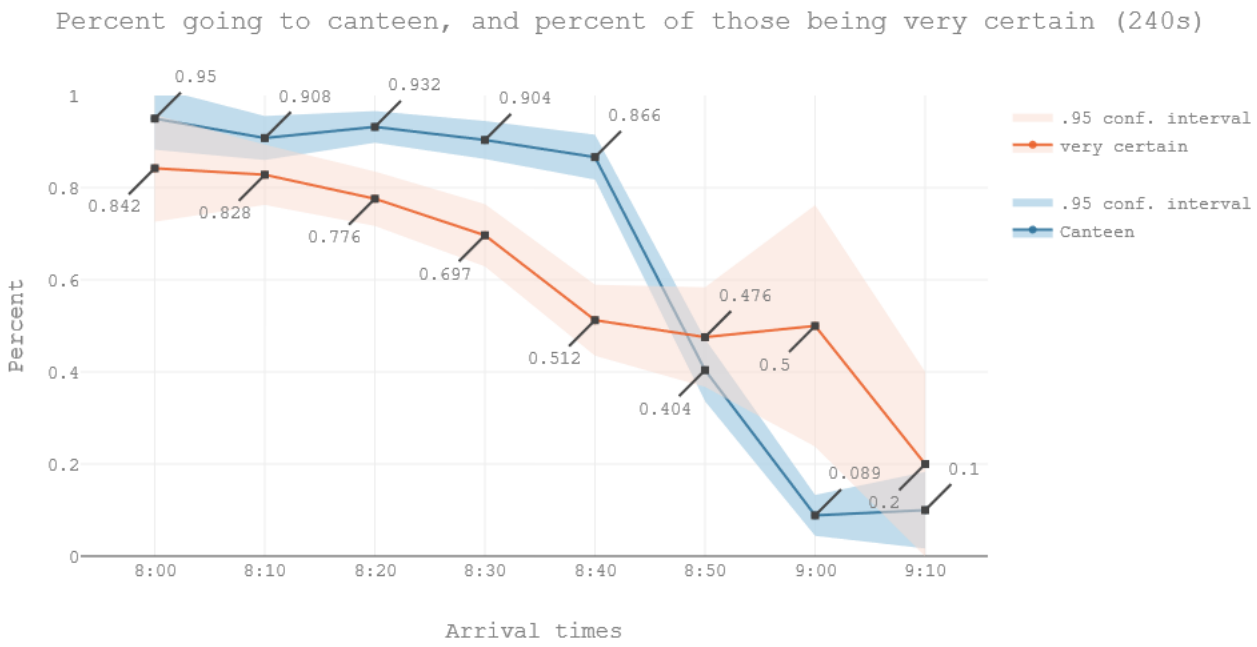
In fact, if a participant arrives at 8:50 and believes for whatever reason that the other player would both choose office at 9:00, and just sometimes at 8:40, it would make office the rational choice at 8:50.

If a participant chooses office at 8:40 or earlier, we might conclude that they have an $n$-order Theory of Mind. If a participant chooses office at 8:40, where they know a canteen-canteen choice would always be the best option, they have to consider not just that the other person might arrive at 8:50, but also that in such a case, the other person would not know that a canteen-canteen choice would be the best option.

We do in fact see 5%-14.4% choosing office at 8:00 to 8:40. It's possible that this is due to Theory of Mind applied by participants. But if we expect that very few participants perform sufficiently higher order social reasoning to go to the office at 8:00 or 8:10, we might not expect such a uniform distribution of office answers between 8:00 and 8:40. In other words, it is likely that a good portion of these office answers are due to random choices as well. This corresponds roughly to the percent of canteen answers at 9:00 and 9:10.

## 1.3 Data visualisation

Plot below shows percent going to canteen (blue trace) and percent of those being very certain (orange trace) + shaded .95 conf. intervals. Only includes treatment with 240s instruction time.



## 1.4 Extra-questions

**Most relevant:**

1. Did you ever go to the canteen at an arrival time later than what is safe according to your previous answer? Why or why not? [Right after 'cutoff' question, free text form]

2. Did you ever make a different decision after seeing the same arrival time again at a later point in the game? Why or why not? [free text form]

**Questions about understanding the rules:**

Questions testing if participants have understood the rules:
(a) you prefer to go to the canteen, but only with your colleague,
(b) you have to go to the office if you or your colleague's arrival time is 9:00 or later, and
(c) you always arrive within 10 minutes of each other.

We might want questions instead that test the belief 'Going to the office together when both players arrive before 9:00, is equally bad as going to different places'.

1. Imagine you arrive at $x$ arrival time. You receive a text from your colleague telling you went/will go to office. Where do you go? (Consider varying $x$ to see if it changes results. Possible answers {canteen, office, does not matter, don't know}, knowledge of (a) entails office is correct choice)

2. Imagine you arrive at 9:00. You receive a text from your colleague telling you he arrived at 8:50. Where do you go? (Possible answers {canteen, office, does not matter, don't know}, knowledge of (b) entails office is only correct choice).

3. Imagine you arrive at 8:50. You receive a text from your colleague telling you he arrived at 9:00. Where do you go (Same as above, office correct answer again).

4. Imagine you arrive at 8:40. You receive a text from your colleague telling you he arrived at 8:50. Where do you go? (Possible answers {canteen, office, does not matter, don't know}, this tests whether participants choosing office more at 8:40 if they know the other arrives at 8:50).

5. Imagine you arrive at 8:20. You know that your colleague always chooses office if they arrive at 8:30. Where do you go? (Possible answers {canteen, office, does not matter, don't know}, this makes 8:20 structurally similar to 8:50, and tests whether participants understood (a) and whether there is a difference in framing the question with an earlier arrival time).

6. When you choose to go to the office before 9:00, e.g. 8:50 and earlier, what was your reasoning behind it?

**Other questions:**

1. What thoughts did you have when making your certainty estimate (Possibly illuminates a hypothesis that there is a disassociation between choices and certainty)

2. Were you sometimes less certain about e.g. going to the canteen? Why did or did you not do that? (Same as above)

3. What do you think about a strategy of only going to the office?