university of groningen / faculty of science and engineering / bernoulli institute

Cooperation and exploitation in a modified Mod game

Artificial Intelligence Bachelor's Project Thesis

Maxime Bos, s3953246, m.r.m.bos@student.rug.nl
Supervisors: mr. J.D. Top, MSc & dr. H. De Weerd

Abstract: Game theory, a branch of mathematics that mainly focuses on the analysis of different strategies in various versions of competitive settings, has been widely researched. Most of this research, however, is performed on the relations between human players. It is interesting to conduct more research on the influence of a human player knowingly playing against a computer, or a simulated agent, on their chosen strategy and presented behaviour versus believing they are playing against another human participant. For the purposes of this research, human participants played a modified version of the Mod Game (an n-player, numerical version of 'rock-paper-scissors'). The modification made to this game is a signalling round in which each player takes turns where they can signal their choice to their opponent. They are allowed to lie, however; it is up to the opponent to decide whether they will trust the player or not and adjust their strategy accordingly. The game was played in multiple different, counterbalanced conditions. In the first condition, players are made to believe they are playing against a human when they are really playing against an agent, whereas in the following conditions, they will knowingly play against an agent. Experimental results show that when humans think they are playing against another human, they tend to trust their opponent less, but they do play more honestly in comparison to when they knowingly play against an agent. The results also show that when an agent plays dishonestly in the very beginning of the block, there is a significant decrease in honesty and trust presented by the participants in the remainder of the blocks. These results suggest that a simulated agent's behaviour, and knowingly playing against an agent rather than a human, does in fact have a significant influence on the strategies chosen and behaviour presented by human participants.

1 Introduction

Game theory is a concept that has been widely researched (Fudenberg and Tirole, 1991; Gibbons et al., 1992; Straffin, 1993). The term 'game theory' is often defined as a branch of mathematics that focuses mainly on analysing different strategies in various versions of competitive settings. The strategies that the players in these situations choose, but also the strategies their *opponents* choose, have to impact the outcome of the game directly (Osborne et al., 2004).

There has been plenty of research on the influence of a human opponent's actions or strategies on the actions chosen by a player (Grosz, Kraus, Talman, Stossel, and Havlin, 2004; Solomon, 1960). Still, it would be very interesting to further research how a human player would respond to playing against a computer or a simulated agent. Would knowing that they are playing against a computer influence these human players' strategies? Would they perhaps be less forgiving and more competitive, or would it have no influence at all?

Especially in competitive settings, social relations are extremely important (McGloin, Hull, and Christensen, 2016). A player could greatly improve their odds of winning either if they pick the right teammate to cooperate with or if they are good enough at deceiving everyone else so they can walk away the lone victor. This player's opponents, in turn, clearly need to be able to know whether the player is being cooperative or deceptive, otherwise they risk losing the game. Knowingly playing against a computer might take away the social connection players feel when they know they are playing against other humans, which could influence the choices they make. It is very possible that they underestimate their opponent if they know they are playing against a computer, which could also influence their chosen actions. It would be interesting to see if this is indeed the case, and if so, what this influence is.

One particularly useful game to test this relation between knowingly playing against a computer or a human and the strategies chosen by human players is through a game known as the Mod Game (Frey, 2013). This game is essentially an n-person extension of rock-paper-scissors. In this research, however, the game will also be played by 2 players, similar to the classic rock-paper-scissors game. First, a range of numbers from 0 to m is chosen. For each round the game is played, the players then simultaneously call out a number from that range. A player gets a point for every time their opponent called out a number that is exactly one lower than the number they called out (with the exception of when the player chose 1, then they get a point if their opponent chooses the highest number (m) in the range). For this research, an extension is added to this game. This extension will be further explained in section 2. As mentioned earlier, if human players play against simulated agents, and these results are compared to games played against human players, it should be possible to find a possible influence of this difference.

This article will investigate the influence of knowingly playing against a simulated agent (versus believing the human players are playing against another human) on the presented behavior by human players in a modified version of the Mod Game.

The following sections will further describe the details of the Mod Game and the modification made to this game for the purposes of this research (section 2), a detailed description of the experiment (section 3), the experiment's preliminary and statistical results (sections 4 and 5) and lastly, a discussion and interpretation of the results, followed by the conclusion of this research (section 6).

2 The Mod Game

For my research I will have the participants play a game known as the "Mod Game" (Frey and Goldstone, 2013). To reiterate, his game is essentially an n-player extension of the classic "rock-paper-scissors" game. However, in this thesis, only the 2-person variant is researched (Veltman, 2017). In this game, a range is chosen between 1 and m, with m $>$ 1. In this thesis, this range is 1 - 24. Every participant has to choose a number in this range. A participant wins the round if they chose a number that is exactly one higher than the number their opponent chose (for example the player chooses the number 4 and the opponent chooses the number 3, then the player gets a point). The only exception to this rule is that the choice of the number 1 always wins over the choice of the number 24. This means that for every choice a participant can make, there is another choice that can beat their choice (just the same as in "rock-paper-scissors"). A Mod Game with two players and a range of 1 to m with $m = 3$ is the same as a non-zero-sum game of rock-paper-scissors.

As mentioned in section 1, in this paper, an extension is added to the original Mod Game. Instead of having both players in the game choose their actions simultaneously, a moment is introduced before each round in which one player communicates which number they will choose to their opponent. The players take turns for who has to declare their choice every round (e.g. before round 1, the player declares their choice and before round 2 the opponent declares their choice and before round 3 it is the player's turn again, et cetera). However, the players do not necessarily have to be truthful. They are allowed to lie about which choice they are going to make in order to try to trick their opponent. This way, they can choose to play cooperatively or deceptively toward their opponent.

When analyzing the responses from the participants in the game, patterns may emerge and the game was originally meant as a method to find those patterns. The Mod Game was originally introduced by Seth Frey and Robert Goldstone (2013). As Frey and Goldstone mention in their article, the Mod Game has a mixed-strategy Nash equilibrium, meaning that there are no players who can increase their expected reward by playing a different strategy when the strategy played by other players remains unchanged. The mixed-strategy Nash equilibrium dictates that each action is chosen with equal probability. However, in case one participant decides to not play according to the randomization strategy, all other players could choose to also play a different strategy in order to try and gain an advantage.

It was found, indeed, that repeated Mod Games elicit behavior from the participants that is not in accordance with the Nash equilibrium (Frey, 2013; Veltman, 2017; de Weerd, Verbrugge, and Verheij, 2014). It would be interesting to see if human players will use predictable behaviour shown by their opponent to try and exploit them for their own gain. It would also be interesting to research whether the players will use a concept known as Theory of mind. Theory of mind is a phenomenon in cognition that describes the ability to attribute thoughts and feelings to others (Premack and Woodruff, 1978). For example, if Alice is throwing a surprise party for Bob, zero-order theory of mind would be that Alice knows about this party. However, if Alice uses first-order theory of mind, she believes that Bob does not know about this party. With second-order theory of mind, Alice would believe that Bob does not know that Alice knows about the party. This can go recursively. Using theory of mind could result in some interesting strategies chosen by the players.

3 Methodology

3.1 Participants

A total of 20 participants (the terms ‘participant’ and ‘user’ will be used interchangeably) played the modified version of the Mod Game. All participants are attending the University of Groningen at the time this research takes place and follow a degree fully conducted in English, meaning they have a good understanding of the language, ensuring they understand the instructions provided.

The participants were all in the age range of 18 to 23 (M = 20), there were 8 female participants and 12 male participants.

All participants are enrolled in a technical BSc or MSc degree (either artificial intelligence, computer science, or similar), ensuring they are familiar with computers and virtual experiment environments. None of the participants, however, had any prior knowledge about the Mod Game.

The experiment has a total duration of about 20 - 35 minutes per participant. The participants were all compensated accordingly for this time.

3.2 Experiment Setup

The participants play the modified Mod Game in three different conditions. In the first condition, they are made to believe through suggestive wording that they are playing against a human when, in reality, they are playing against a simulated agent (refer to section 3.3.2 for an in-depth explanation of this agent’s behavior). In the last two conditions, the participants know that they are playing against an agent. However, they are not told what strategy the agent is using. A more in-depth explanation of the agent strategy implementation for each condition is given in section 3.3.2. The game environment and experiment setup will be explained in this section.

3.2.1 Starting information

Before the experiment begins, the participants receive an informed consent form that briefly explains the experiment’s goals. After they sign this form, the experiment begins. First, the participants are shown a starting screen. This screen briefly explains the Mod Game, how it works, and what the rules are. Specifically, it tells the participants what is expected of them. This explanation also includes an explanation of the extension I have added to the original Mod Game, in which each participant has to tell their opponent which number they will play in the next round (in this case, it’s not used to tell others what they should play, which may be an interesting variation for further research, see section 6). They are explicitly told they can either tell the truth and therefore cooperate with their opponent or lie to try and trick their opponent. They are also told that they can change their strategy at any time (meaning if they choose to cooperate in the first round, they do not have to cooperate for the entire game. They can choose to play dishonestly in other rounds). Lastly, the explanation informs the participants that there is no time limit. They can take as long as they want to think of a choice to make. Before each condition, a separate and different start screen is shown:

After the main start screen with the explanation of the Mod Game, the participants get two trial rounds in order to become familiar with the system and the game. After these trial rounds, the participants play three different conditions. For all participants, the first condition is the same, but the next two conditions are counter-balanced (see section 3.3.2).

For the first condition, the participants are made to believe that they are playing against a human opponent while they are actually playing against a simulated agent (see section 3.3.2). This is done by having two participants do the experiment simultaneously in the same room (divided by a short wall, so they cannot see each other’s screen) and using specific wording that suggests each participant is playing against the other. If only one participant shows up, or only one participant registered to a specific timeslot, a confederate (an ‘actor’) is used to obtain the same condition. The start screen for this condition leads the participants to believe that they are playing against a human opponent.

For the second condition, the participants are told they are playing against a simulated agent. They are not told, however, what strategy this agent will play.

For the third condition, the participants are shown the same start screen as in the second condition. The only difference is that it is stated very clearly that the agent they are playing against for this condition is a different agent than the one they previously played against. This should prevent the participants from having any negative feelings towards the agent, which could influence their strategy choice.

The agent and participant scores are shown on the screen during the game rounds. However, as opposed to the original implementation of the Mod Game model (Veltman, de Weerd, and Verbrugge, 2019), when a round is over, it says “player/opponent gained a point” instead of “player/opponent won this round” at the top of the screen. If rounds are “won”, it may prime participants for competition and prevent cooperation. This is to be avoided since this research is focused on examining the behavior presented by the participants based on the agent's behavior, rather than on any outside factors.

3.2.2 Total duration of experiment

Each participant plays 20 rounds per condition, with the signalling phase before each round. The participant and agent take turns with who signals per round and the agent always starts. There is a short break in between each condition. Every participant plays for approximately 20-35 minutes to an hour, including short breaks.

In the break between the first and second condition and between the second and third condition, the participants are given a short questionnaire with questions such as "What do you think of the level of difficulty of the game?", "Briefly describe your strategy, if you had any" et cetera (questionnaires can be found in appendix E). This is meant to "reset" the participants for the next condition and should also prevent the participants from getting tired or bored.

After finishing the third condition, the participants get a p-beauty contest (López, 2001). This is a test where participants are shown a range of numbers from 0 to 100 and are asked to choose the number they think will be chosen the most on average divided by 2 (i.e. if they think 50 would be the number chosen the most on average, they will select 25). The winner of the game is the person who chooses the number closest to the average of all numbers chosen by the participants (p-beauty test can be found in appendix E). The p-beauty contest is very useful to determine which order of Theory of Mind* a participant has to try and win the game. It is interesting to see if the order of ToM the participants used has any influence on whether or not they won the game. Perhaps further research could be performed on this effect (see section 6 for future research).

3.3 JavaScript model basis

For this experiment, I used the general JavaScript implementation of the Mod Game used in the research by Veltman, De Weerd and Verbrugge (2019) about training the use of Theory of Mind using artificial agents. In this implementation, there is already a model for the Mod Game that I used as a basis for my experiment. However, I did implement some alterations in the original code to make the model conform to what my experiment looks like. This section will explain the alterations I implemented and an overview of the simulated agents' different strategies in each condition.

3.3.1 Model alteration

The main alteration made to the original JavaScript model is the addition of the signalling round. This is the moment before each round in which a player has to inform their opponent of which number they are going to choose in the following round (the signal). They can either lie and signal a different number than they are actually going to play to try and trick their opponent, or they can tell the truth in an attempt to work together. The players take turns with who can signal for each round. The number that is signalled turns red so the opponent can see clearly which number was selected. When the agents present honest and trusting or dishonest and distrusting behaviour differs between each condition (refer to section 3.3.2 for further explanation). The participant's signal is retrieved by user input, whereas the agent signals a random number ranging from 1 to 24.

3.3.2 Agent behaviour choices

Before getting to the details of the agent behaviour imlplementations, some terms should be explained. Two concepts mainly present in the following sections are 'trust' and 'honesty'.

Trust can be defined as the player choosing the signal + 1 whenever the opponent signalled (denoted with the symbol 'T') (and distrust when they do not, denoted with the symbol 'D'). Honesty can be defined as the player choosing the signal whenever they themselves signalled (denoted with the symbol 'H') (and dishonest when they do not, denoted with the symbol 'X').

Trial rounds: For the two trial rounds that are presented to the participant before the actual game starts, the agent plays only honestly and trusting. Honesty can be defined as the agent choosing the number they signalled, and trust is defined by the agent choosing the participant's signal + 1. This way, the participants are not primed for dishonesty, but they do get a good understanding of how the game works and what the agent's behaviour can look like.

In all conditions, the agent starts signalling. For the first 10 participants, the agent's behaviour in each of the conditions is as follows:

Condition 1: For condition 1, the human participants are led to believe they are playing against a human instead of an agent. Therefore the agent's behaviour needs to be less predictable than in the following two conditions.

*the ability to attribute mental states to others, such as beliefs, desires, intentions, goals, et cetera (de Weerd et al., 2014) the participants experienced during the experiment. Humans often use this ability in social settings to understand better the situation they're in.

The agent's behaviour in this condition is based on the actions and strategies chosen by the human participants. In the first round, the agent signals their choice first. In this round, the agent always plays honestly (meaning they choose the number they signalled). For all other rounds, the agent's behaviour is a reaction to the behaviour presented by the participants.

First, a 'cooperation probability' is calculated. This cooperation probability is then compared to a randomly generated number (between 0 and 1). In case the random number is lower than the probability that was calculated, the agent plays honestly and trusting in the subsequent round. Otherwise, the agent plays dishonestly and distrusting. The way this cooperation probability is calculated is as follows: First, the amount of honestly/trustingly played rounds in the last three rounds are counted. This number is then divided by 3 (the 3 previously played rounds, both agent and participant signal rounds) to determine what the probability is that the agent plays honestly and trusting in the current round. If it is determined that the agent will play honestly and trusting, they will play the number they signalled in case it was the agent's turn to signal, or the agent will play the number that is one higher than the signal in case it was the participant's turn to signal in the current round. However, if it is determined that the agent will play dishonestly and distrusting, then the algorithm is a bit more complicated. First, it is calculated how many consecutive rounds the agent has not been able to gain a point. In case the agent has not been able to gain a point for three consecutive rounds, it cheats. This is implemented to make the agent more realistic by behaving less predictable. Will they start playing more dishonestly or distrusting? Or will their behaviour not change at all? It is important to note that the way this cheating behaviour is implemented should be realistic enough that the participants cannot tell the agent is cheating. The agent cheats by waiting for the participant to make a choice and then choosing a number that is one higher than that choice, ensuring the agent gains a point. However, the agent can only do this if the participant's choice is less than 10 numbers higher than the signal and if the participant's choice is not lower than the signal, otherwise it would be too obvious that the agent cheated, which might cause the participants to become suspicious. If the agent has been able to gain a point within the last three rounds, the agent's choice is calculated differently. This choice is determined using a so-called 'change rate'. This change rate is calculated by determining what the honest or trusting choice by the participant in the last round should have been. This honest or trusting choice would have been the signal in all rounds where the participant signalled or the signal + 1 in all rounds where the agent signalled. The change rate is calculated by finding |actual choice - honest choice| + 1. Once this change rate is calculated, the algorithm looks at the current round. the agent's final choice is then determined by calculating the change rate + cooperative choice % 24.

In this condition, it is very important that the human participant believes they are playing against another human for the entire game duration. Therefore I also implemented a minimum length of round duration. This means that a round has to take at least N seconds. When the human participant chooses which number to play in the round, it is calculated how many milliseconds they are short of or over this duration limit (the same goes for the signalling phase). If they responded in under N seconds, the result of the round would not be shown until after the N seconds have passed. If they take longer than N seconds, the results are shown immediately. For the signalling phase, when the agent has to signal, a random duration is selected from within a range based on the average time it took the human participant to choose a signal. This choice duration is determined using a pilot study, where one participant plays the game and their reaction time is measured between each action an agent takes and each choice the participant makes for both signalling and making a final choice of which number to play. This gives the illusion that the human participant's opponent has to "think" about their choice, which makes it more believable that it is not, in fact, a computer simulation. Both the duration for the choice lag and the signal lag have been determined using the average duration presented by a human participant in the pilot study. For the very first round of the experiment, the timeout was set at 5000 ms to make it look like the opponent needs to get used to the game environment.

The choice for response lag N was determined after conducting the pilot study and measuring the average response time of the participant. According to the pilot data, the participant had an average response time of around 5000 ms. When the agent plays honestly and distrusting, the timeout is anywhere between 2500 and 3000 ms since the agent should not have to think as long to make a choice as in comparison to when it is playing deceptively. This further increases the believability of the agent being a human player. When the agent plays dishonestly and distrusting, the timeout is anywhere between 4500 and 5500 ms. A window was chosen rather than one set number of seconds because it is more realistic when a round has differing lengths rather than one set time duration.

Condition 2.1: For condition 2, the agent plays honestly and trusting for all twenty rounds. The agent does this by choosing the signal + 1 whenever the participant signals or by playing the signal whenever the agent signals. Having the agent play only this way is interesting to see whether the human participant will go along with this and thus also play honestly and trusting, so both parties get a lot of points, or if the participant will be dishonest to try and get the most points for himself (and thus throw the agent under the bus so to speak).

Condition 3.1: For the last condition, the agent starts by immediately playing dishonestly in the first round. The agent does this by playing the agent's signal + 2. After the first round, the agent plays only honestly and trusting for the next fourteen rounds. The last five rounds are played dishonestly and distrusting. The agent's behaviour for these last five rounds is hardcoded as follows: for round 16, the agent cheats; for round 17, the agent plays the agent's signal + 2; for round 18, the agent cheats again; and for round 19, the agent plays a choice that ensures neither player gains a point by choosing the signal + 3. With this condition, it becomes possible to see whether the human participant plays honestly anyway or if they immediately distrust the agent and start playing dishonestly in a bid to get ahead. If they start playing dishonestly, how long does it take for the human participant to play honestly again, if they even do at all? And if they play honestly after a few rounds of the agent playing honestly, will this change when the agent starts playing dishonestly again?

For the second 10 participants, the first condition remains the same. However, the agent's behaviour in the second and third conditions differ:

Condition 2.2: For this condition, the agent plays the first round dishonestly. This time the agent does this by cheating (see the explanation for condition 1). All remaining rounds are played honestly and trusting.

Condition 3.2: For this condition, the agent plays the first fifteen rounds honestly and trusting and the last five rounds dishonestly and distrusting. The agent's behaviour for these last five rounds was hardcoded in the same way as in condition 3.1.

By counterbalancing the last two conditions for all participants, I should be able to properly investigate the influence of the agent's behaviour on the participant's behaviour.

3.4 What is measured

For every condition, I measured a number of different variables to be able to properly determine the influence of simulated agents' behaviour on the elicitation of honesty or dishonesty in human participants. The variables I measured are the following:

3.4.1 Honesty levels from participants

This is the most important variable. This variable shows how much honest and dishonest behaviour the participants showed their opponents in each of the three conditions. This variable can be easily measured by tracking which choices the participants said they would make every time it was their turn to communicate their choice to their opponent (i.e. the signals). When these signals are compared to the choices they made, a value can be calculated that shows exactly how often the participants were deceptive or cooperative to their opponent. When the signal and the actual choice were equal, the participants played honestly, but when the signal differed from the actual choice, the participants played dishonestly.

3.4.2 Trust levels from participants

This variable is similar to the previous one, only instead of measuring the amount of honest or dishonest behaviour the participant showed their opponent, it is measured how many times the participant decided to trust their opponent. This can be measured by tracking the choice the opponent said they were going to make (i.e. the opponent's signal) and comparing this to the choice the participant made (e.g. if the agent said they were going to choose three and the participant chooses the number 4, it means they trusted the agent. If the participant chose another number, it could mean they distrusted the agent).

3.4.3 P-beauty contest

In addition to all variables mentioned above, the levels of ToM of all participants were measured by the p-beauty contest the participants did after finishing the last condition.

3.4.4 Comparison

Once all these variables are measured and determined they can be compared across all three conditions. Do human participants trust (what they believe to be) other humans more than they trust agents? Are they more honest or dishonest depending on who they think their opponent is?

It is important to note that in all conditions, the opponent is actually an agent (even when the human participants think they are playing against another human). This is important since then the three conditions can be properly compared without having to take other aspects of human behaviour into consideration that may be used/displayed by an actual human opponent. The interpretation of the results is given in the next section.

4 Results - Surface Level Behaviour

The following two sections describe the results of the experiment described in section 3 in order to determine whether (and how) the agent's behaviour influences the behaviour of the human participants.

Before any statistical analyses are performed, it is essential to look at the raw data. This section will describe the preliminary results that are visible in multiple figures depicting the raw data. Section 5 will then describe the statistical analyses performed on this data with regard to trust, honesty and other aspects of the participants' behaviour during the experiment.

4.1 Distribution of signals

Figure 4.1 shows the distribution of the users' signals in block 1 (fake human condition). This figure shows that the distribution of signals seems fairly random, with a preference towards signalling the number 1. The data also showed (not pictured in the figure) that most participants chose to signal the number 1 in the first round of the block where they were allowed to signal. Interestingly, for all other conditions, the distribution of signals is also seemingly random, with a slight preference towards signalling the number 1 but also for some higher numbers (7 and 13 mostly). Figures A.1 - A.4 in appendix A show the distributions of user signals over conditions 2.1 to 3.2.

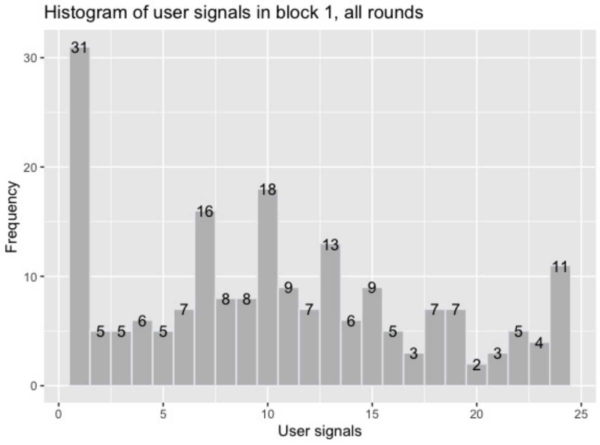


Figure 4.1: Distribution of user signals over all rounds in block 1 (Fake Human condition)

4.2 Distribution of choices

Figure 4.2 shows a figure similar to the one discussed above, except this figure shows the distribution of user choices, rather than signals, over all rounds in block 1. What stands out here is that similarly to the signal distribution, the choices are also seemingly random, with again, a slight preference towards the number 1. However, the choices are more evenly distributed, which I believe is due to the fact that the choices the participants eventually make are based on the provided signals.

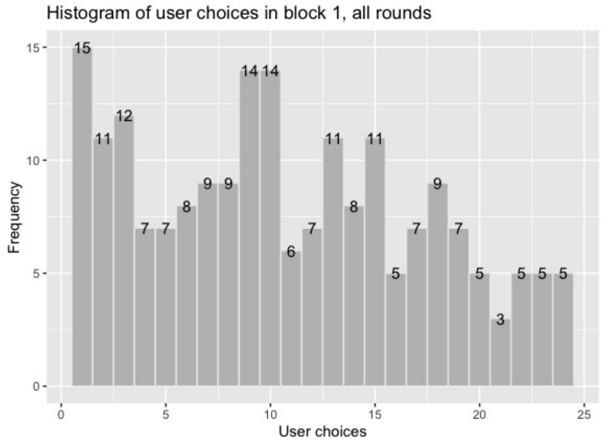


Figure 4.2: Distribution of user choices over all rounds in block 1 (Fake Human condition)

For blocks 2.1, 2.2, 3.1 and 3.2, the distribution of choices is similar to the distribution of choices in block 1: choices are seemingly random, with some slight preferences. Most of these preferences correspond to a preferred signal + 2. For example, for all agent blocks, user signals peaked at 1, and user choices peaked at 3, correspondingly. For brevity, the figures showing the trends for these blocks are omitted but can be found in Appendix B.

4.3 Distribution of user choices relative to agent signals

Figure 4.3 shows the distribution of the user choices relative to the agent's signals over all rounds in block 1 (meaning user choice - agent signal). What is interesting here is that there seems to be an apparent preference for choosing a number that is one higher than the agent signal, which could indicate that the participants trust that their opponent will choose the number they signaled (for statistical analyses, refer to section 5.1). Choosing a number that is one higher than the signal should, according to their belief, result in the participant getting a point. However, as you can see in the figure, there is a slight peak at -1. The participants decided 10 times to choose the agent signal - 1, which would cause *the agent* to gain a point rather than themselves. This is a fascinating form of cooperation. Looking at the data, there was 1 participant who chose this form of cooperation in the first few rounds but switched to choosing the agent signal + 1 after a while. The other times the agent signal - 1 was chosen, multiple different participants chose it only once during arbitrary rounds, so it seems these occurrences were not consistent attempts at cooperating by playing agent signal - 1.

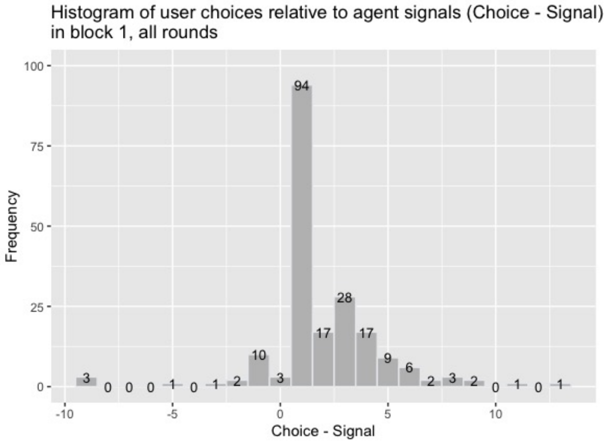


Figure 4.3: Distribution of user choices relative to agent signals over all rounds in block 1 (Fake Human condition)

Figure 4.4 shows the distribution of user choices relative to agent signals for block 2.1 (20 x Honest/Trusting). The behaviour presented in this figure shows that there is a clear preference towards choosing the agent signal + 1. Since this agent's behaviour was very predictable, the participants seemed to have tried to gain as many points as possible. Again, we see one instance where a participant chose the agent signal - 1. This participant had not yet done this (also not in block 1), and it was in the middle of the game, so it is fair to assume these occurrences were also not consistent attempts at cooperating by playing agent signal - 1. There are also small peaks at the signal + 2 and + 3. These choices all occurred at the beginning of the game, suggesting the participants were distrustful of the agent and were trying to think one step ahead to gain a point. They quickly adapted as the game went on and proceeded to play the signal + 1. Figures C.1 - C.3 show the distributions for all other conditions. Especially for the conditions where the agent played deceptively for a period of time, there are also peaks at the signal + 2, which suggests the participants quickly adapted to the agent's deceptive play.

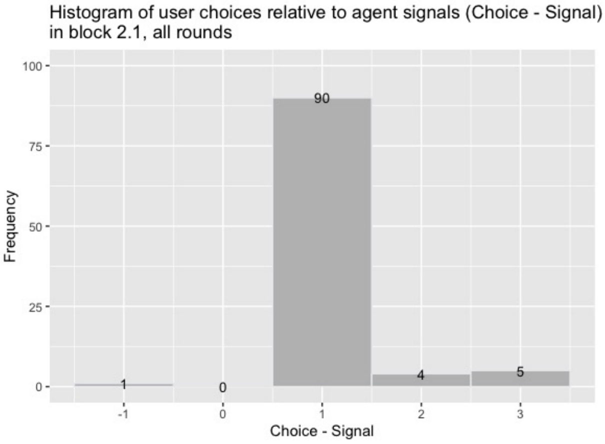


Figure 4.4: Distribution of user choices relative to agent signals over all rounds in block 2.1 (20 x Honest/Trusting)

4.4 Distribution of user choices relative to user signals

The distribution of user choices relative to user signals is very interesting. Figure 4.5 shows this distribution for block 1 over all rounds. It is clear that the participants showed a preference for choosing the number they signaled. There is also a prevalent appearance of participants who decided to choose the signal + 2 or even + 4 seemingly to try and gain a point for themselves. The peak at the number 0 is interesting because it suggests that these participants decided to play trusting and honestly, thus allowing the opponent to gain a point. When looking at the data, however, it does show that most participants started playing honestly and trusting in this case but switched their strategies once the opponent started playing more dishonestly. The peaks at 0, +2, +4, and +6 suggest participants had different levels of the theory of mind concept (refer back to 2 for further explanation). In the case of the results showing the user choice relative to the user signal, the peaks suggest which order of theory of mind the participants used. When the peak is at +2, the participants may believe that the agent knows they are not going to choose the signal, therefore it would choose a number one higher than the signal. The participant thus chooses a number one higher than that number to try and stay one step ahead of the agent. This could indicate first-order theory of mind. The peaks at +4 and +6 suggest these participants use even higher orders of theory of mind to rationalize their eventual choices.

Figure 4.6 shows the distribution of the user choice relative to the user signal for block 2.1 (1 x Dishonest, 19 x Honest/Trusting). This block has a clear preference for choosing the number 2 higher than the participants' signal. The partici-

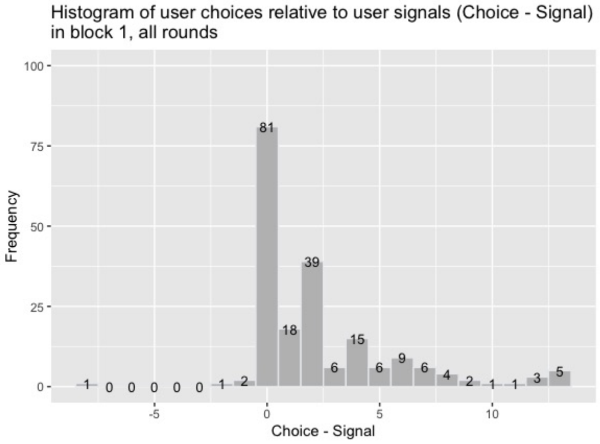


Figure 4.5: Distribution of user choice relative to user signal over all rounds in block 1 (Fake Human condition)

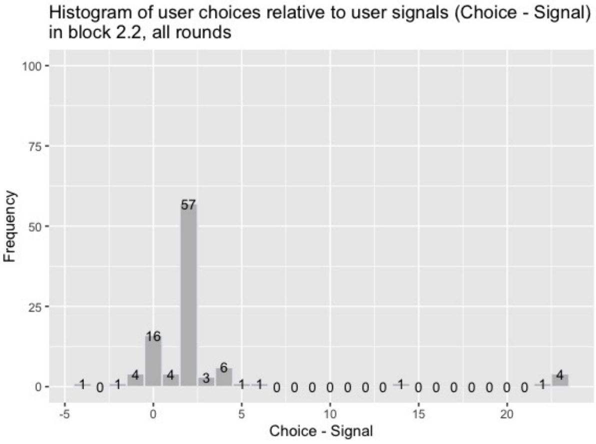


Figure 4.6: Distribution of user choice relative to user signal over all rounds in block 2.2 (1 x Dishonest, 19 x Honest/Trusting)

pants seem less likely to play honestly towards the agent and are more apprehensive about having the agent gain a point. It appears they would rather get points for themselves. This is an important distinction between this block and the distribution for block 1 because, in this block, the participants know they are playing against an agent when they were led to believe they were playing against a human opponent in the first block. Were they more lenient in block 1 perhaps because they thought they were playing against a human?

The distributions for all other conditions show similar trends and are therefore not included in this section. Please refer to Appendix C for all figures of these distributions. The data from these blocks show the behaviour that was predicted when designing the experiment. When the agent played (mostly) honestly, the users chose to gain points for themselves rather than allowing the agent to gain a point. They often chose either the signal they gave + 2 or + 4 depending on how often the agent played dishonestly in each block. This behaviour is also reflected in the distribution for the user scores, as shown in figure 4.7. This figure clearly shows that participants received much higher scores in the agent blocks, as opposed to the first (fake human) block. This distribution seems to suggest that when an agent started playing a round dishonestly (blocks 2.1 and 3.2), the participants trusted the agent less, causing them to choose the user signal + 2 or the agent signal + 3 more often, even when the agent would play honestly the rest of the block.

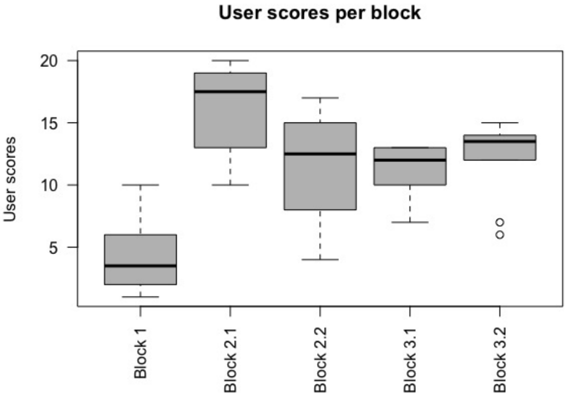


Figure 4.7: Boxplot of distribution of scores all users achieved during all blocks

4.5 P-beauty and questionnaire results

Unfortunately, the p-beauty contest did not result in any interesting or useable insights as it seemed from the results that the participants did not properly understand the task, and will therefore not be discussed in section 5. It is interesting to note that 85% of participants referred to the other participant in the room in the questionnaires when talking about what they think their opponent's strategies were in the first block. This suggests they truly believed they were playing against a human opponent. This is good because it means that the results discussed in the following sections are valid.

5 Results - Statistical Analyses

5.1 Trust levels

Figure 5.2 shows the trust levels of each participant for all the conditions described in section 3.3.2. A trust level can be defined as the percentage of the number of times a participant decided to trust that the agent would actually choose the number they signalled. For example, if the agent signalled the number 3, the participant can 'trust' the agent and believe they are actually going to choose the number that was signalled. A participant who trusts the agent will therefore choose the signalled number plus one (the number 4 in this case) to gain a point. The agent can signal ten times in each block. The following analyses are performed in order to investigate the influences of the agent's behavior on the choices the participants make. For the analyses of block 1 compared to block 2 and 3 combined, specifically, it is investigated what the influence is of the participants believing they are playing against a human rather than an agent.

5.1.1 Block 2.1 vs block 2.2

Figure 5.2 shows the trust levels for conditions 2.1 and 2.2, respectively. Block 2.1 represents the first ten participants playing the game according to condition 2.1 described in section 3.3.2. Here, the agent played all twenty rounds, both trusting and honest (meaning they played the signal they gave and chose the signal + 1 whenever the participant signalled). Block 2.2 represents the second ten participants playing the game according to condition 2.2 described in the last part of section 3.3.2. The agent played all rounds trusting and honest as well, except for the first round, where it played dishonestly (meaning it chose a number different from the signal it gave in a bid to 'trick' the opponent). It is interesting to investigate whether the chosen type of action of the agent in that first round influences the trust levels in the participants for the rest of the block.

When I compare the trust levels for block 2.1 (M = 90, SD = 18.86) and block 2.2 (M = 63, SD = 17.03), the result of two-proportions chi-squared test indicates that the type of action chosen by the agent in the first round (X or H) has a significant effect on the trust levels of the participants in the remainder of the block ($\chi^2(1)$ = 14.61, p = 0.00013)). A two-proportions chi-squared test was chosen over a two-tailed, unpaired t-test because the data was skewed, and a t-test would not have yielded a valid result. For the remainder of the trust level analyses a proportion test will also be used.

5.1.2 Block 3.1 vs block 3.2

Figure 5.2 also shows the trust levels for blocks 3.1 and 3.2. Block 3.1 represents the first ten participants playing condition 3.1, as described in section 3.3.2, whereas block 3.2 describes condition 3.2 for the second ten participants, as stated in section 3.3.2. In block 3.1, the agent plays round 1 dishonestly, rounds 2 - 15 trusting and honest (agent signal, or participant signal + 1), and rounds 16 - 20 dishonest again. In block 3.2, the agent plays the same way as in block 3.1, except for round 1, where the agent plays trusting and honestly instead of dishonestly. As described in the subsection above, I investigated the influence of the agent's choice of action in this first round on the trust levels of all participants in these blocks.

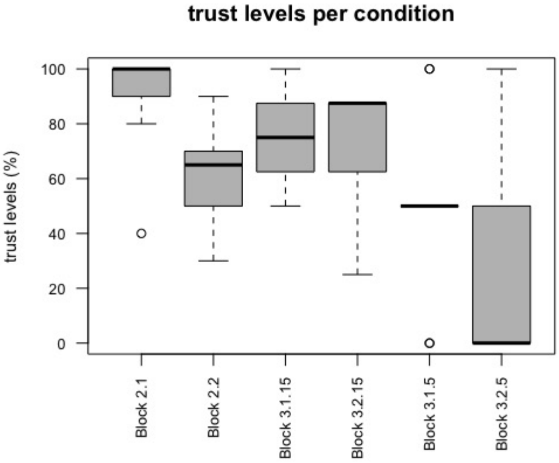


Figure 5.1: Trust levels for conditions 2.1 [20 x H/T], 2.2 [1 x X, 19 x H/T], first 15 rounds of 3.1 [1 x X, 14 x H/T, 5 x D/X], first 15 rounds of 3.2 [14 x H/T, 5 x D/X], last 5 rounds of 3.1 and last 5 rounds of 3.2. Only the rounds where the agent signaled were used for calculation of percentages.

After comparing the trust levels for block 3.1 (M = 71, SD = 15.95) and block 3.2 (M = 63, SD = 20.03), the result of the two-proportions, chi-squared test suggests that for these blocks, as opposed to blocks 2.1 and 2.2, the agent's choice of type of action in this first round is not of significant influence on the participants' trust levels ($\chi^2(1)$ = 0.78, p = 0.378).

What is interesting is that between only the first 15 rounds of block 3.1 (M = 76.25, SD = 14.97) and the first 15 rounds of block 3.2 (M = 72.5, SD = 21.89), a two-proportions, chi-squared test also showed no significant difference ($\chi^2(1)$ = 1.07e-30, p = 0.99). See figure 5.1 for the trust levels for the first 15 rounds of conditions 3.1 and 3.2

Similarly, between only the last 5 rounds of blocks 3.1 (M = 50, SD = 33.3) and 3.2 (M = 25, SD = 35.36), a two-proportions, chi-squared test showed, again, no significant difference ($\chi^2(1)$ = 2.74, p = 0.097). Figure 5.1 shows the trust levels for the first 15 rounds of both conditions 3.1 and 3.2 and the last 5 rounds of both conditions 3.1 and 3.2.

5.1.3 Block 1 vs block 2 + block 3

Figure 5.2 show the trust levels of all participants in block 1, block 2 (conditions 2.1 and 2.2 combined) and block 3 (conditions 3.1 and 3.2 combined). The trust levels of block 1 are compared with those of block 2 and block 3 combined to see whether there is a significant difference in trust levels when participants think they are playing against a human compared to when they think they are playing against an agent. The results of comparing block 1 (M = 47, SD = 32.2) with the combination of blocks 2 and 3 (M = 71.75, SD = 17.035) show that there is indeed a significant difference between the participants believing they are playing against a human in comparison to when they know they are playing against an agent. This comparison was performed using a two-proportions, chi-squared test ($\chi^2(1) = 21.85$, p = 2.94e-06).

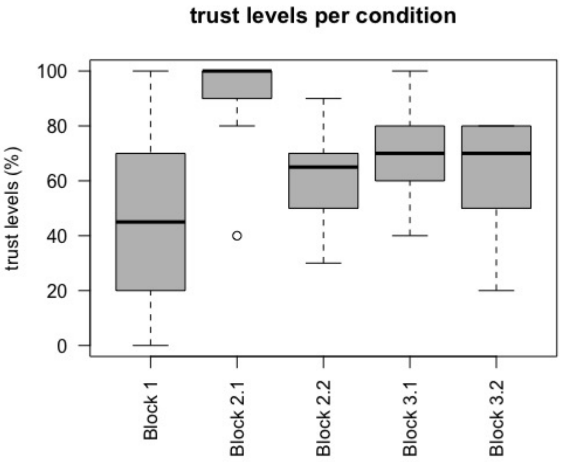


Figure 5.2: Trust levels per participant group over all rounds of conditions 1 (Fake Human), 2.1 (20 x H/T), 2.2 (1 x X, 19 x H/T), 3.1 (1 x X, 14 x H/T, 5 x D/X) and 3.2 (15 x H/T, 5 x D/X). Only the rounds where the agent signalled were used for calculation of percentages.

5.1.4 Order effects

In order to check whether there are order effects between blocks 2 and 3, there are a few more comparisons that need to be analyzed.

First, it needs to be determined whether there is a significant difference between the trust levels of block 2.1 (M = 90, SD = 18.86) and the first 15 rounds of block 3.2 (M = 72.5, SD = 21.89). Both of these conditions are all played honestly and trusting by the agent, so if there is a significant difference between the trust levels, it could mean that there are indeed order effects. A two-proportions, chi-squared test, however, showed no significant difference between these two conditions ($\chi^2(1) = 2.93$, p = 0.087). Figure 5.1 shows the trust levels for condition 2.1 and the first 15 rounds of condition 3.2.

Second, the difference between condition 2.2 (M = 63, SD = 17.03) and the first 15 rounds of condition 3.1 (M = 76.25, SD = 14.97) needs to be analyzed. These two conditions both consist of the first round being played dishonestly by the agent, while the subsequent rounds are all played honestly and trusting. A two-proportions, chi-squared test showed that between these conditions, like in condition 2.1 compared to the first 15 rounds of 3.2, there is no significant difference ($\chi^2(1) = 3.79$, p = 0.0514). Figure 5.1 shows the trust levels for condition 2.2 and the first 15 rounds of condition 3.1.

5.2 Honesty levels per participant

Figure 5.3 shows a boxplot of the percentage of honestly played rounds per participant for all blocks. To repeat, an honest play is defined by the player choosing the signal whenever they signalled. An honest play could therefore be seen as a round in which the player allows the opponent to gain a point. Similarly to subsection 5.1, the aim here is to investigate the influence of the agent's behavior on the behaviors the participants show. It is also important to investigate the influence of the belief that the participants are playing against a human rather than against an agent. These analyses are described in the following sections. Similarly as with the trust level analyses, all honesty level analyses are tested using a two-proportions, chi-squared test because the data was heavily skewed and a t-test would not have yielded valid results.

5.2.1 Block 2.1 vs block 2.2

Figure 5.3 shows the honesty levels for blocks 2.1 and 2.2. Block 2.1 represents the first ten participants playing condition 2.1, as described in section 3.3.2, whereas block 2.2 describes condition 2.2 for the second ten participants, as stated in section 3.3.2. In block 2.1, the agent plays all rounds honestly and trusting (meaning it plays agent signal and user signal + 1 for all rounds). In block 2.2, the agent plays the same way as in block 2.1, except for round 1, where the agent plays honestly instead of dishonestly. As described in the subsection above, I investigated the influence of the agent's choice of action in this first round on the trust levels of all participants in these blocks.

After comparing the honesty levels for block 2.1 (M = 19, SD = 31.43) and block 2.2 (M = 16, SD = 22.71), the result of the two-proportions, chi-squared test suggests that for these blocks, the agent's choice of type of action in this first round is not of significant influence on the participants' trust levels ($\chi^2(1) = 0.139$, p = 0.71).

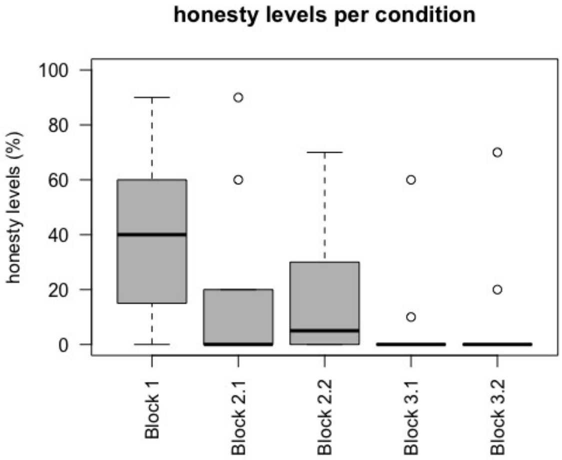


Figure 5.3: Honesty levels per participant group over all rounds of conditions 1 (Fake Human), 2.1 (20 x H/T), 2.2 (1 x X, 19 x H/T), 3.1 (1 x X, 14 x H/T, 5 x D/X) and 3.2 (15 x H/T, 5 x D/X). Only the rounds where the user signalled were used for calculation of percentages

5.2.2 Block 3.1 vs block 3.2

Figure 5.3 also shows the honesty levels for blocks 3.1 and 3.2. Block 3.1 represents the first ten participants playing condition 3.1, as described in section 3.3.2, whereas block 3.2 describes condition 3.2 for the second ten participants, as stated in section 3.3.2. In block 3.1, the agent plays round 1 dishonestly, rounds 2 - 15 trusting and honest (agent signal, or participant signal + 1), and rounds 16 - 20 dishonest and distrusting again. In block 3.2, the agent plays the same way as in block 3.1, except for round 1, where the agent plays honestly instead of dishonestly. I, again, investigated the influence of the agent's choice of action in this first round on the honesty levels of all participants in these blocks.

After comparing the honesty levels for block 3.1 (M = 7, SD = 18.89) and block 3.2 (M = 9, SD = 22.34), the result of the two-proportions, chi-squared test suggests that for these blocks, the agent's choice of type of action in this first round is not of significant influence on the participants' honesty levels ($\chi^2(1) = 0.068$, p = 0.794).

Like in section 5.1.2, between only the first 15 rounds of block 3.1 (M = 10, SD = 26.98) and the first 15 rounds of block 3.2 (M = 8.58, SD = 22.54), a two-proportion, chi-squared test also showed no significant difference ($\chi^2(1) = 0.01$, p = 0.99).

Again, the results of a two-proportion, chi-squared test showed no significant difference between only the last 5 rounds of blocks 3.1 (M = 0, SD = 0) and 3.2 (M = 10, SD = 22.498) ($\chi^2(1) = 1.404$, p = 0.236). Figure 5.4 shows the honesty levels for the first 15 rounds of both conditions 3.1 and 3.2 and the last 5 rounds of both conditions 3.1 and 3.2.

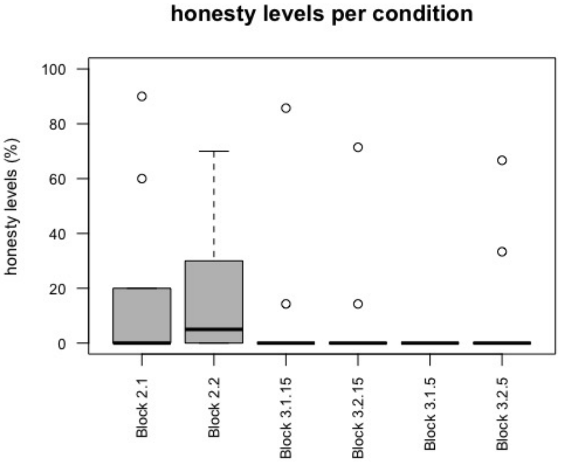


Figure 5.4: Honesty levels for conditions 2.1 [20 x H/T] , 2.2 [1 x X, 19 x H/T], first 15 rounds of 3.1 [1 x X, 14 x H/T, 5 x D/X], first 15 rounds of 3.2 [14 x H/T, 5 x D/X], last 5 rounds of 3.1 and last 5 rounds of 3.2. Only the rounds where the user signalled were used for calculation of percentages

5.2.3 Block 1 vs block 2 + block 3

Figure 5.3 shows the honesty levels of all participants in block 1, block 2 (conditions 2.1 and 2.2 combined) and block 3 (conditions 3.1 and 3.2 combined). The honesty levels of block 1 are compared with those of blocks 2 and 3 combined to see whether there is a significant difference in honesty levels when participants think they are playing against a human compared to when they think they are playing against an agent.

The results for block 1 (M = 40.5, SD = 28) and blocks 2 + 3 (M = 12.75, SD = 19.02) indicate that there is indeed a significant influence ($\chi^2(1) = 58.22$, p = 2.34e-14).

5.2.4 Order effects

Like in section 5.1.4, the order effects also need to be analyzed here. The same conditions are compared.

First, it needs to be determined whether there is a significant difference between the honesty levels of block 2.1 (M = 19, SD = 31.43) and the first 15 rounds of block 3.2 (M = 8.58, SD = 22.54). Both of these conditions are all played honestly and trusting by the agent, so if there is a significant difference between the honesty levels, it could mean that there are indeed order effects. A two-proportion, chi-squared test showed, however, that there is not a significant difference between these two conditions ($\chi^2(1) = 2.79$, p = 0.095). Figure 5.4 shows the trust levels for condition 2.1 and the first 15 rounds of condition 3.2.

Second, the difference between condition 2.2 (M = 16, SD = 22.71) and the first 15 rounds of condition 3.1 (M = 8.57, SD = 22.54) needs to be analyzed. These two conditions both consist of the first round being played dishonestly by the agent, while the subsequent rounds are all played honestly and trusting. A two-proportion, chi-squared test showed that, as in condition 2.1 compared to the first 15 rounds of 3.2, there is no significant difference ($\chi^2(1) = 0.806$, p = 0.369). Figure 5.4 shows the trust levels for condition 2.2 and the first 15 rounds of condition 3.1.

5.3 Average reaction times

It is also worth researching whether the agent's actions influence the participants' reaction time. When an agent did something unexpected in the previous round, you expect the reaction time to be higher in the subsequent round since the participants might have to take more time to think about how to respond. The reaction time, in this case, is defined as the difference in time between the signal appearing and the participant making their action choice.

Figure 5.6 shows the logarithm of the reaction times per participant over all rounds where the agent signalled in block 1, where the previous round was played honestly by the agent (HRT) and where the previous round was played dishonestly by the agent (XRT). Exploratory analysis showed that the reaction times appear to be skewed. Because of this, the logarithms of the reaction times were taken and used in the statistical analyses as well as in figure 5.6. Comparing the average reaction times (using a paired, two-sample t-test) when the previous round was played honestly (M = 8.197, SD = 0.466) with the average reaction times when the previous round was played dishonestly (M = 8.232, SD = 0.383) yields no significant difference (t(19) = -0.303, p = 0.765). This suggests that the type of choice of the agent does not influence the participants' reaction times.

5.4 User scores per block

When visually inspecting figure 4.7, which shows the total score for all participants per block, it seems that the scores for blocks 2.1, 2.2, 3.1 and 3.2 are much higher than those for blocks 1. Since the agent behaves much more predictably in blocks 2 and 3 than in block 1, I investigated whether there is a significant difference in participant scores

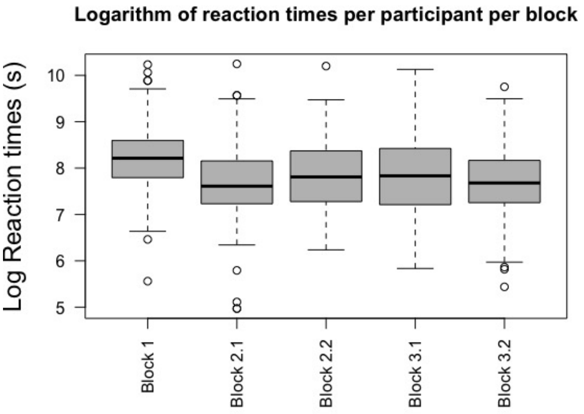


Figure 5.5: Logarithm of reaction times for all participants over all rounds and all blocks.

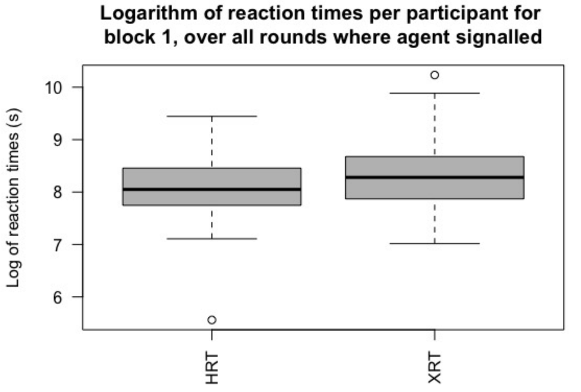


Figure 5.6: Logarithms of reaction times after either an honestly (HRT) or dishonestly (XRT) played round by the agent per participant over all rounds in block 1 (Fake human condition)

between these blocks. When comparing the scores from block 1 (M = 4.15, SD = 2.72) to blocks 2 and 3 combined (M = 12.65, SD = 2.85), the two-sample, dependent t-test shows a significant difference in user scores (t(19) = -10.14, p = 4.2e-09) which suggests that the participants do in fact seem to take advantage of the agent's predictable behavior to gain more points for themselves.

6 Discussion & conclusion

When interpreting the results, some interesting trends come to light. Starting with the results of section 4: The histograms of choices and signals made by the participants show that these distributions seem fairly random, with a slight preference towards choosing the numbers 1, 3, 9, and 10 for the choices, and the numbers 1, 7, and 10 for the signals. While the preferences were for slightly different numbers than the ones mention in the article by Veltman (2017), this is still in line with the results found there. More interesting, however, are the distributions of user choices relative to the user and agent signals. The combination of the distributions of trust and honesty can show four different scenarios:

Scenario 1: high levels of trust combined with high levels of honesty can point to the participant trying to cooperate with the agent to ensure both parties receive points.

Scenario 2: high levels of trust combined with low levels of honesty can point to the participant being exploitative, meaning the participant tries to gain as many points for themselves as possible.

Scenario 3: low levels of trust with low levels of honesty. This scenario suggests the participant is competitive. They try to gain as many points as possible for themselves and do not trust their opponent to give them honest signals, so they choose other numbers with the belief it will provide them with points.

Scenario 4: low levels of trust combined with high levels of honesty. A possible implication of this scenario could be that the participant is sacrificing their own point gain for that of their opponent. A possible explanation for this happening could be that the participant believes that they are playing against someone special to them (a friend, a partner, et cetera), and they do not want to hurt their feelings.

The experiment results focused on these trust and honesty levels show that, in general, the participants showed behaviour that is in accordance with both scenarios 1 and 3 for the fake human block.

Interestingly, however, almost all participants showed extremely low levels of honesty and higher levels of trust in blocks 2.1 and 2.2 (extremely high levels of trust in block 2.1 and with the levels of trust in block 2.2 slightly lower than those in block 2.1). Since the agent's behaviour in these blocks is very predictable, it suggests that the participants are playing very exploitatively (scenario 2). One participant played a few rounds according to scenario 4 in block 1, and even they played very exploitatively in blocks 2.1 and 2.2. According to section 5.1, there is a significant difference between the trust levels of blocks 2.1 and 2.2, which shows that the dishonest behaviour of the agent in the first round greatly influences the behaviour of the participants in the following rounds. For block 2.1, the participants play very exploitatively, whereas, for block 2.2, the participants play more competitively.

For blocks 3.1 and 3.2, there is an even lower level of honesty shown in almost all participants for blocks 3.1 and 3.2 than in comparison to the levels shown for blocks 2.1 and 2.2. The trust levels for both conditions are similar to those of condition 2.2, with no significant difference between the trust levels of blocks 3.1 and 3.2. This is interesting because while the first round played dishonestly by the agent in block 2.2 had a significant influence on the behaviour of the participants, it does not seem to influence the participants in block 3. These lower trust levels in block 3 (both conditions) could also be attributed to the fact that the agent played the last 5 rounds of these conditions dishonestly, causing the participants to trust the agent less.

When comparing the trust and honesty levels between block 1 and the combinations of all results of blocks 2 and 3 (all conditions), it becomes clear that there is a significant difference in trust and honesty levels for all participants. This suggests that the participants' belief they are playing against a human rather than an agent does indeed significantly influence the participants' behaviour.

6.1 Improvements

Even though a lot of preparation and commitment went into forming the experiment performed for this research, there are of course always unforeseen troubles that might influence the results. This section will discuss some possible improvements that could further improve the validity of the results.

One of such improvements is for example that in this research, block 1 (the fake human condition) was always played first. After this block 2 and block 3 would follow, respectively. Since all participants did the experiment in this manner, there is no way to tell whether there are order effects between these blocks that might possibly influence the results.

Another issue that arose during the statistical analyses is that it is not clear from the results whether the differences between block 1 and blocks 2 and 3 were due to the participants believing they were playing against a human rather than an agent, or because the agent's behaviour was more predictable in the last two blocks. Similarly, it is now hard to distinguish whether certain behaviour was presented due to the participant thinking they are playing against a human, to them not fully understanding the game yet, or to simply playing against a more complex agent than in the other blocks.

Another improvement that could be made to improve the validity of the results further is to explicitly ask the participants after the experiment is completed whether they believed they were playing against a human participant in block 1. However, when looking at the answers written in the questionnaires, 35% of participants referred to the participant they thought they were playing against directly when answering a question about the strategy they believed their opponent was playing in block 1. But, since this is not the majority of the participants, it is unclear whether the rest truly believed they were playing against the other person in the room. For future research, therefore, it is better to explicitly add this extra question at the end of the experiment.

6.2 Future research

It is important to note that while research on human behaviour and intelligence has been performed using the Mod Game before (de Weerd, Verbrugge, and Verheij, 2013; Veltman et al., 2019; Veltman, 2017; de Weerd et al., 2014), no research had been done yet using this game with the signalling phase that was added in this research. This is a new tactic to investigate game strategy and human behaviour. In prior research involving the Mod Game, it has only been used as a competitive game. This research is the first that utilizes the Mod Game as a method to investigate human cooperative behaviour. There are multiple possibilities to use this version of the Mod Game in further research.

In this research, the participants are explicitly told to use the signalling phase as a way to signal the choice they are going to make to their opponent. It is unknown how the participants would interpret the signalling phase in case they are not given these explicit instructions. Will they still use it to signal their own choice? Or will they signal the number they want or expect their opponent to choose? Or would they simply signal a random number, not realizing they can use this signalling phase to their advantage if used well?

Another interesting variation is to have the participants actually play against other humans (rather than a 'fake human') and compare these results to the fake human agent from this experiment. This would be useful to improve the validity of the results further. Of course it is also possible, and very interesting, to research whether the results found in this experiment still hold when the participants play the fake human condition twice, with the only difference being that the second time they are told they are playing against a simulated agent. Another option is to have one group of participants knowingly playing against an agent, and the other group playing against what they believe to be a human, and then comparing the results. This could prevent order effects. Repeating the experiment in this way would prove whether simply knowing the player is playing against an agent significantly influences their chosen strategies and behaviour. Finally, since the analysis of the order effects between conditions 2 and 3, between the first 15 rounds of conditions 3.1 and 3.2 and between the last 5 rounds of conditions 3.1 and 3.2 revealed that there are no differences and thus no order effects, it is likely that there were also no order effects between the fake human condition and the other blocks. However, it is still important that for any future research, a better counterbalancing method is found to make sure there are indeed no order effects.

Conclusion

This research provides evidence that a human player thinking they are playing against another human results in more cooperative and competitive behaviour compared to thinking they are playing against an agent since it's evidenced that agents that behave more predictably and which are perceived as agents and not as humans are more often exploited by human players.

Since these results, detailed in section 5, show that there are indeed significant differences in the behaviour shown by the participants as a response to different agent behaviours, it can be concluded that the answer to the question of whether a simulated agent's behaviour can significantly influence a human participant's behaviour in a modified Mod Game is, in fact, yes.

7 Acknowledgements

I would like to express my special thanks to my supervisors mr. J.D. Top and dr. H. de Weerd, for supporting and guiding me through the process of preparing, researching, and eventually writing this article.

I would also like to thank my confederate, mr. Jord van Es, for his help during the experiment.

Finally, I want to thank the University of Groningen for providing me with an office to conduct my experiment and for the funds that allowed me to compensate my participants for their time.

References

Harmen de Weerd, Rineke Verbrugge, and Bart Verheij. How Much Does It Help To Know What She Knows You Know? An Agent-Based Simulation Study. *Artificial Intelligence*, 199:67–92, 2013.

Harmen de Weerd, Rineke Verbrugge, and Bart Verheij. Theory Of Mind In The Mod Game: An Agent-Based Model Of Strategic Reasoning. In *ECSI*, pages 128–136, 2014.

Seth Frey. *Complex Collective Dynamics In Human Higher-Level Reasoning; A Study Over Multiple Methods*. PhD thesis, Indiana University, 2013.

Seth Frey and Robert L Goldstone. Cyclic Game Dynamics Driven By Iterated Reasoning. *PloS one*, 8(2):e56416, 2013.

Drew Fudenberg and Jean Tirole. *Game Theory*. MIT press, 1991.

Robert Gibbons et al. *A Primer In Game Theory*. Harvester Wheatsheaf New York, 1992.

Barbara Grosz, Sarit Kraus, Shavit Talman, Boaz Stossel, and Moti Havlin. The Influence Of Social Dependencies On Decision-Making: Initial Investigations With A New Game. *http://nrs.harvard.edu/urn-3:HUL.InstRepos:2640578*, 2004.

Rafael López. On P-Beauty Contest Integer Games. *UPF Economics and Business Working Paper*, (608), 2001.

Rory McGloin, Kyle S Hull, and John L Christensen. The Social Implications Of Casual Online Gaming: Examining The Effects Of Competitive Setting And Performance Outcome On Player Perceptions. *Computers in Human Behavior*, 59: 173–181, 2016.

Martin J Osborne et al. *An Introduction To Game Theory*, volume 3. Oxford University Press New York, 2004.

David Premack and Guy Woodruff. Does The Chimpanzee Have A Theory Of Mind? *Behavioral and Brain Sciences*, 1(4):515–526, 1978.

Leonard Solomon. The Influence Of Some Types Of Power Relationships And Game Strategies Upon The Development Of Interpersonal Trust. *The Journal of Abnormal and Social Psychology*, 61(2):223, 1960.

Philip D Straffin. *Game Theory And Strategy*, volume 36. MAA, 1993.

de Weerd Harmen Verbrugge Rineke Veltman, Kim. Socially Smart Software Agents Entice People To Use Higher-Order Theory Of Mind In The Mod Game. In *The 29th Benelux Conference on Artificial Intelligence*, 2017.

Kim Veltman, Harmen de Weerd, and Rineke Verbrugge. Training The Use Of Theory Of Mind Using Artificial Agents. *Journal on Multimodal User Interfaces*, 13(1):3–18, 2019.

A Distribution of user signals

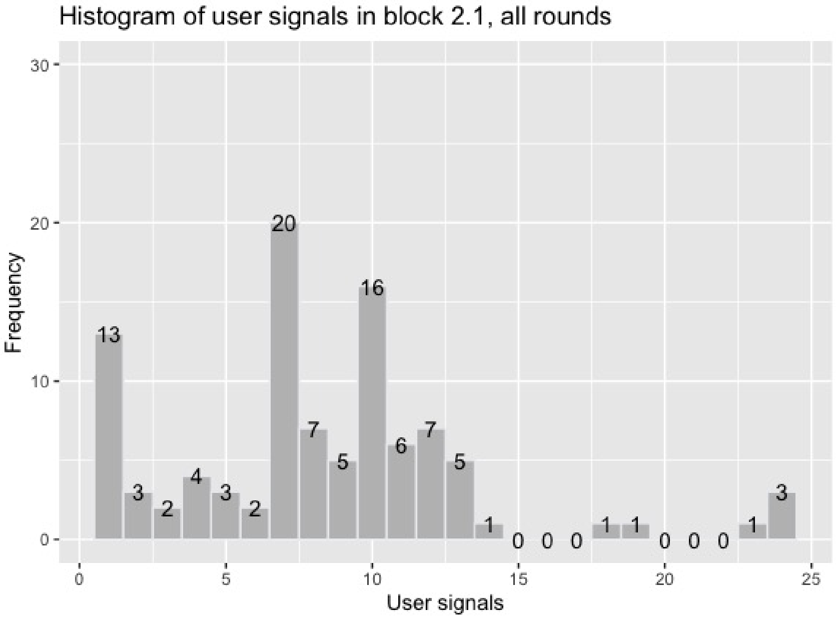


Figure A.1: Distribution of user signals over all rounds in block 2.1 (20 x H/T)

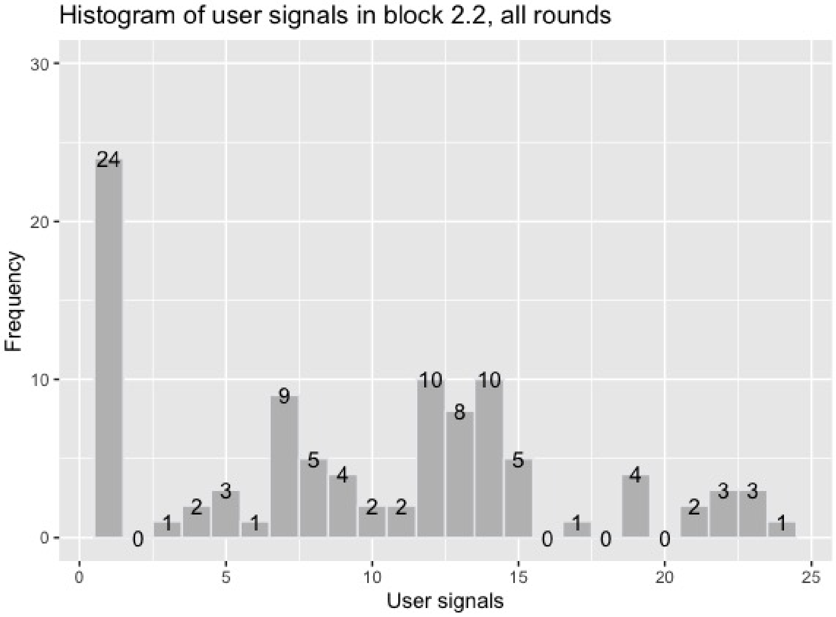


Figure A.2: Distribution of user signals over all rounds in block 2.2 (1 x X, 19 x H/T)

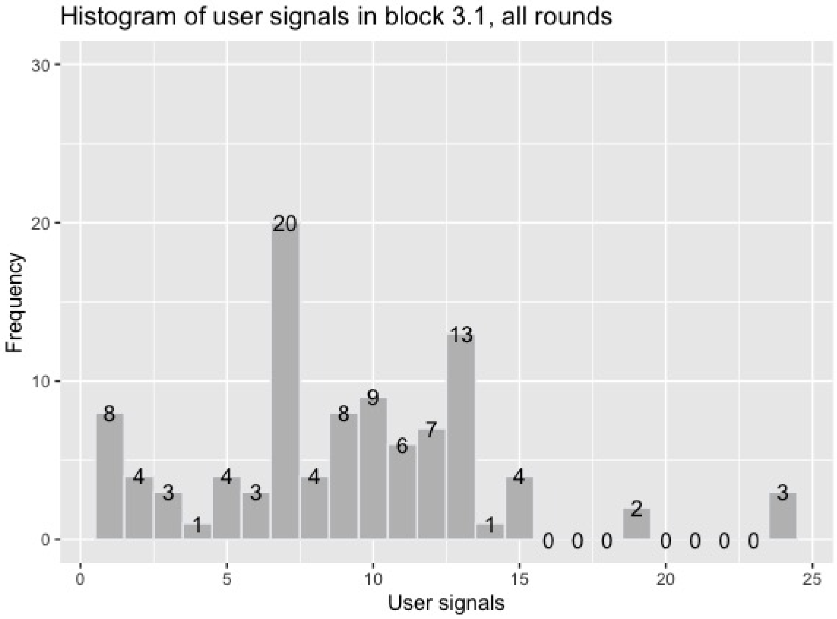


Figure A.3: Distribution of user signals over all rounds in block 3.1 (1 x X, 14 x H/T, 5 x D/X)

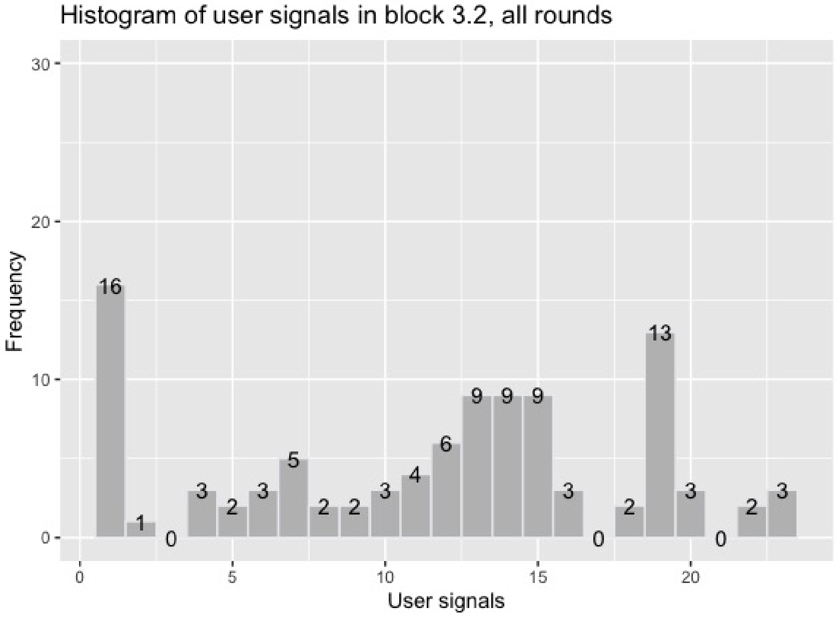


Figure A.4: Distribution of user signals over all rounds in block 3.2 (15 x H/T, 5 x D/X)

B Distribution of user choices

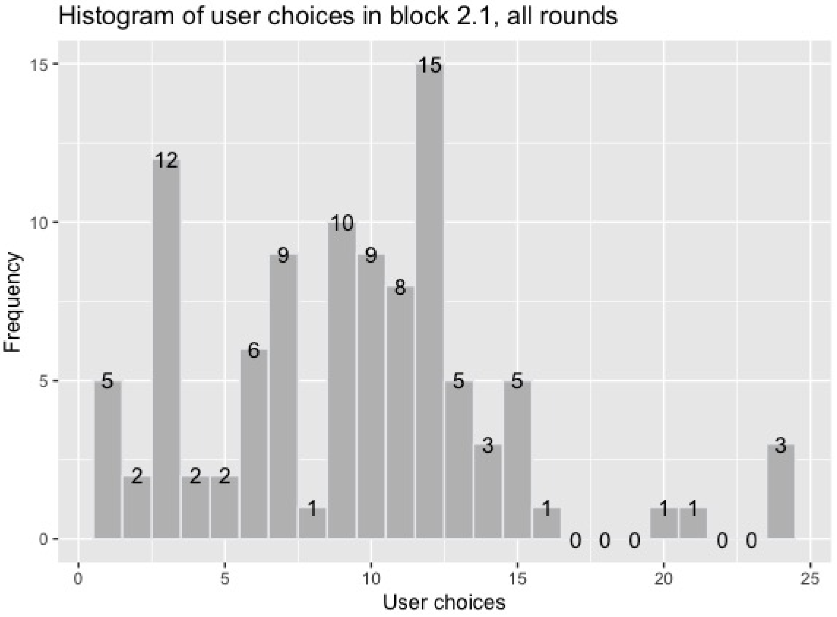


Figure B.1: Distribution of user choices over all rounds in block 2.1 (20 x H/T)

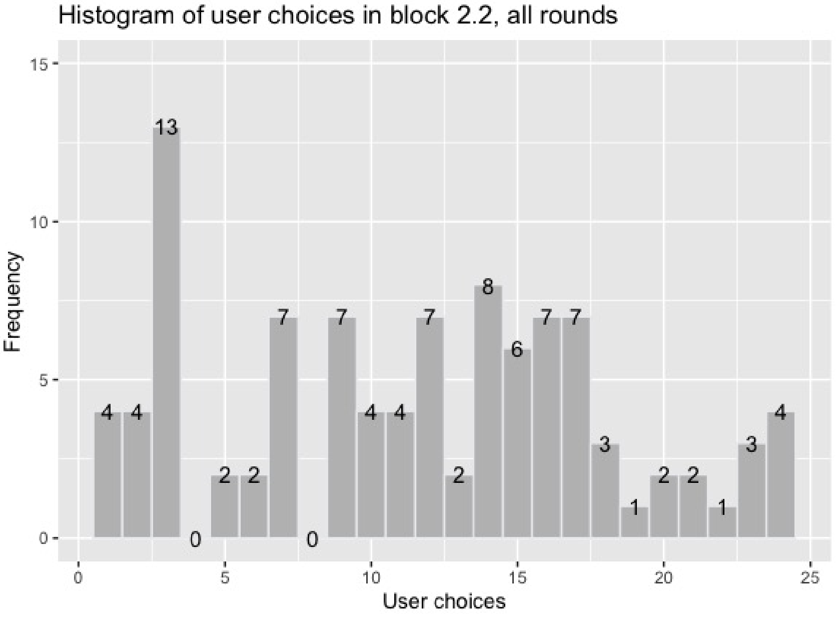


Figure B.2: Distribution of user choices over all rounds in block 2.2 (1 x X, 19 x H/T)

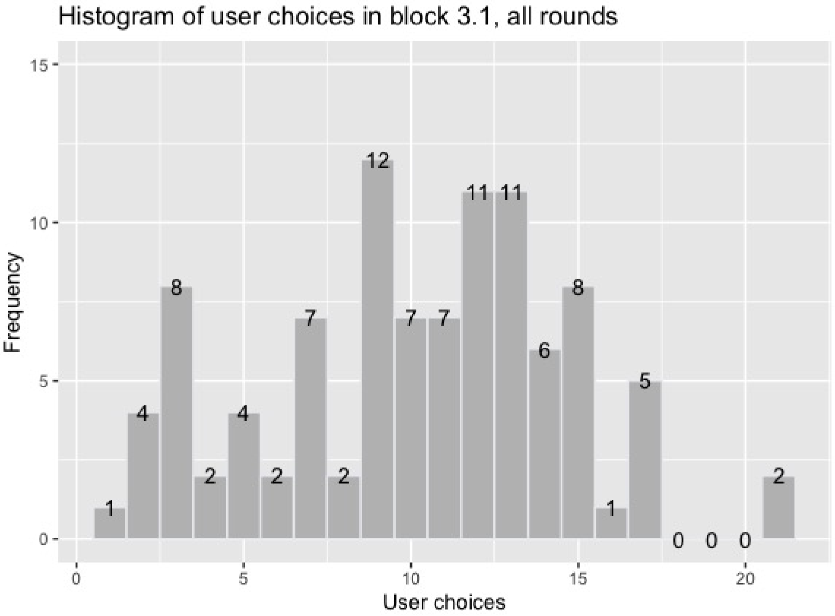


Figure B.3: Distribution of user choices over all rounds in block 3.1 (1 x X, 14 x H/T, 5 x D/X)

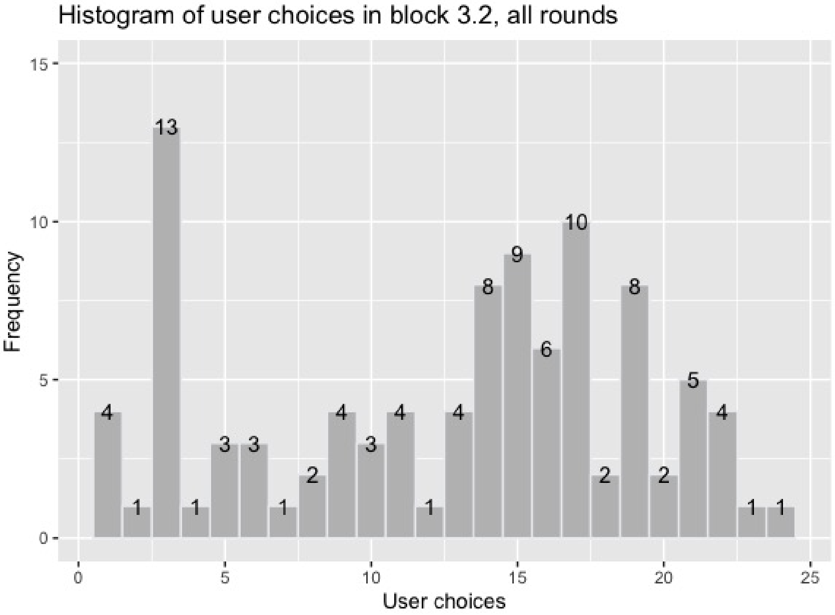


Figure B.4: Distribution of user choices over all rounds in block 3.2 (15 x H/T, 5 x D/X)

C Distribution of user choices relative to agent signals

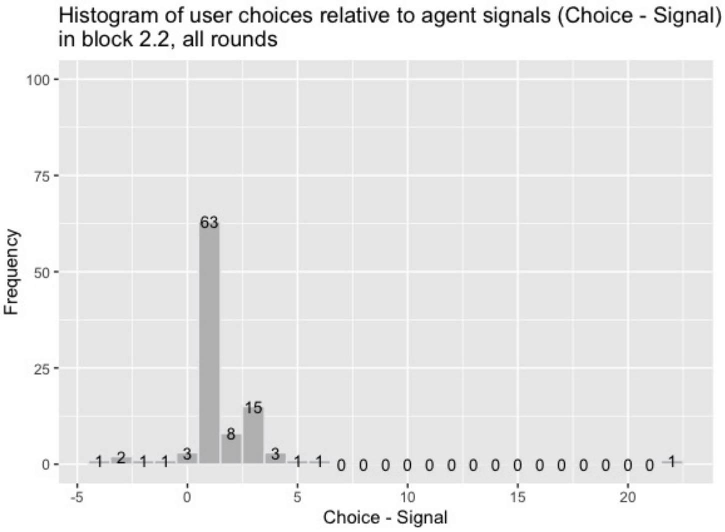

Histogram of user choices relative to agent signals (Choice - Signal) in block 2.2, all rounds

Figure C.1: Distribution of user choices relative to agent signals over all rounds in block 2.2 (1 x X, 19 x H/T)

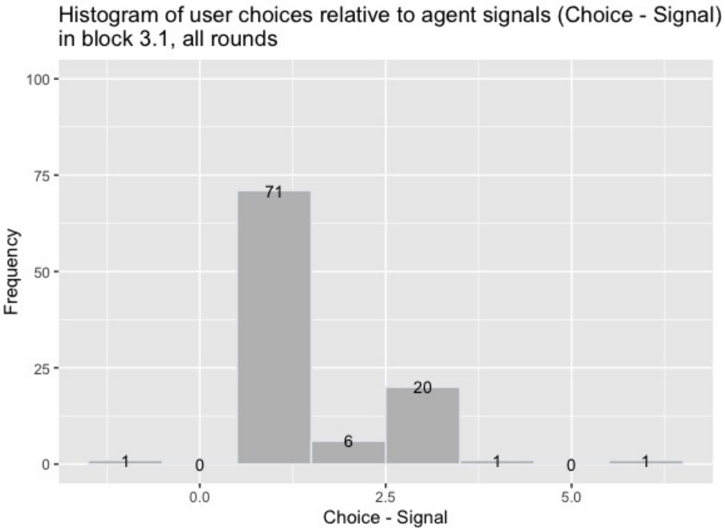

Histogram of user choices relative to agent signals (Choice - Signal) in block 3.1, all rounds

Figure C.2: Distribution of user choices relative to agent signals over all rounds in block 3.1 (1 x X, 14 x H/T, 5 x D/X)

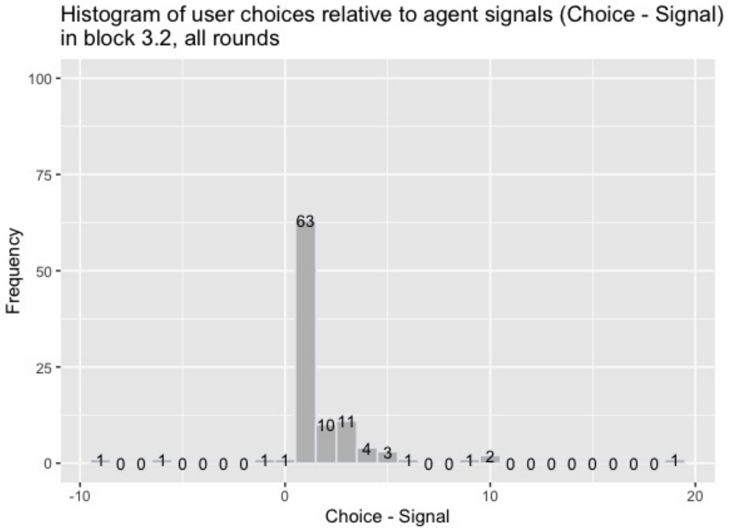


Figure C.3: Distribution of user choices relative to agent signals over all rounds in block 3.2 (14 x H/T, 5 x D/X)

D Distribution of user choices relative to user signals

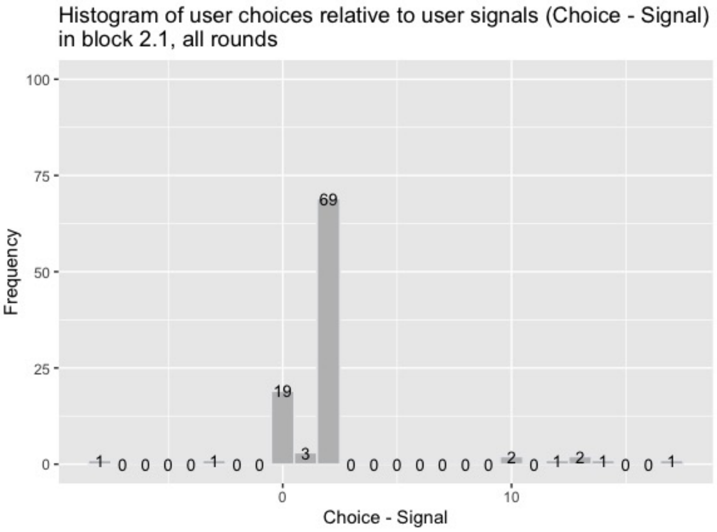


Figure D.1: Distribution of user choices relative to user signals over all rounds in block 2.1 (20 x H/T)

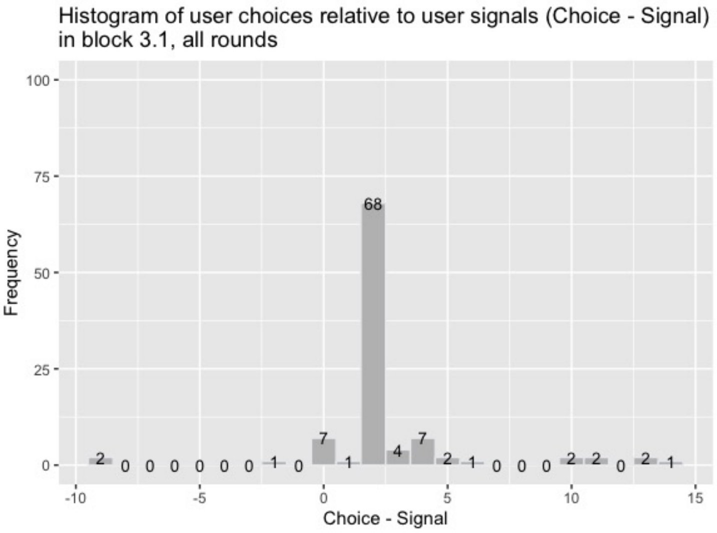


Figure D.2: Distribution of user choices relative to user signals over all rounds in block 3.1 (1 x X, 14 x H/T, 5 x D/X)

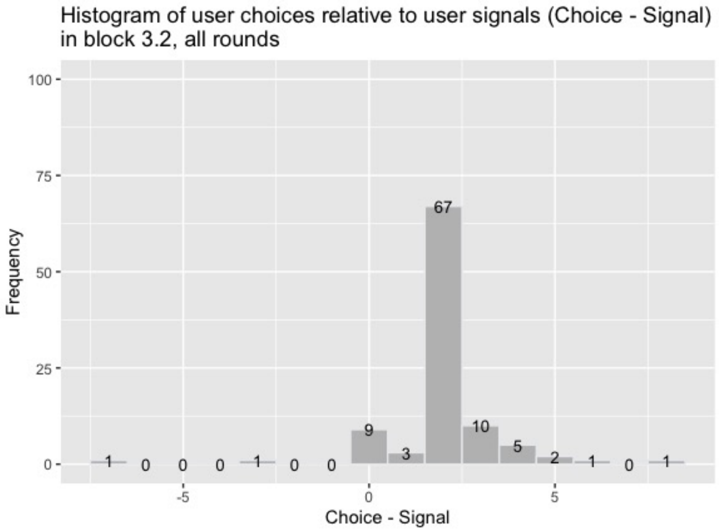
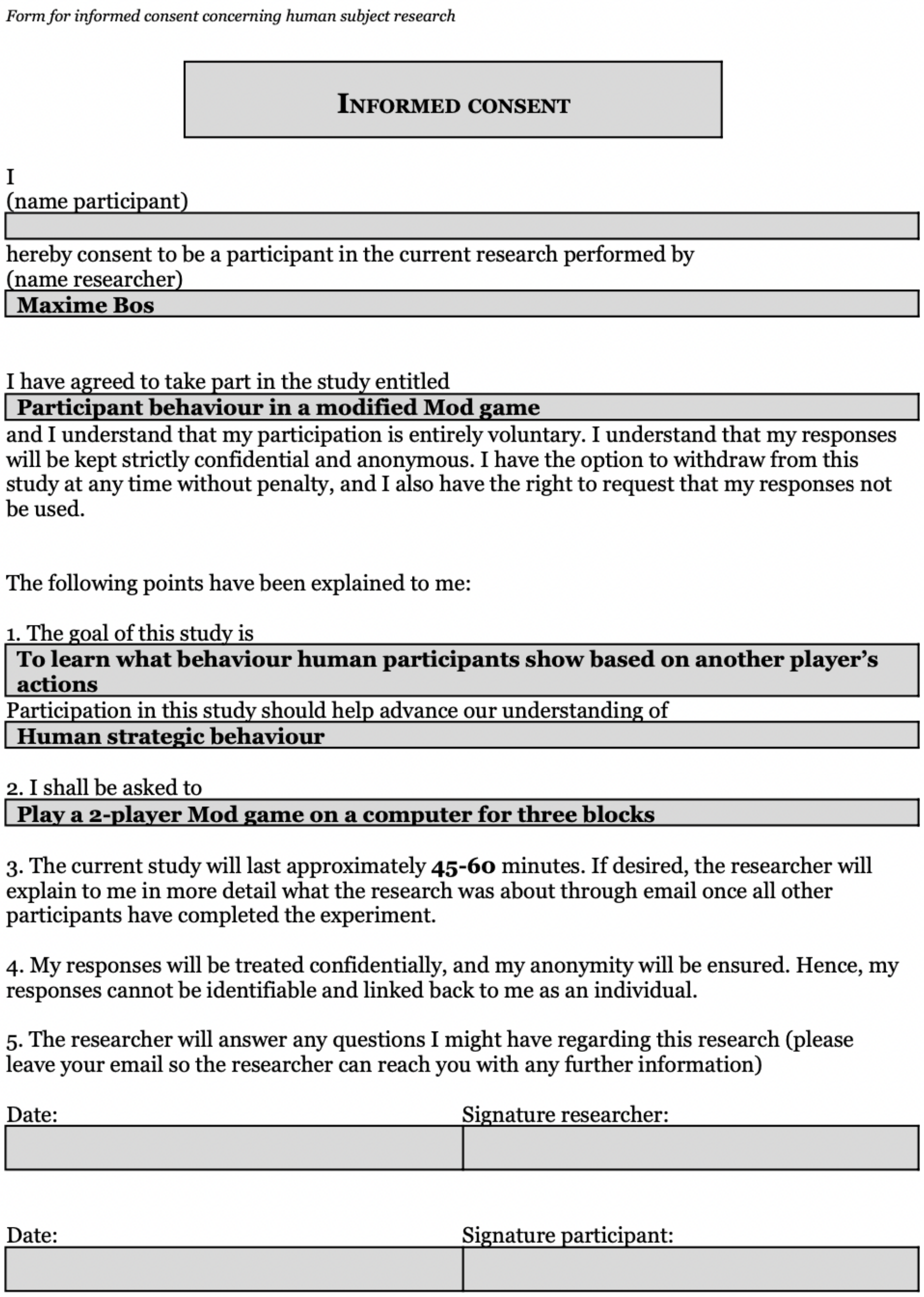


Figure D.3: Distribution of user choices relative to user signals over all rounds in block 3.2 (15 x H/T, 5 x D/X)

E Supporting documents



Form for informed consent concerning human subject research

Informed consent

I
(name participant)

hereby consent to be a participant in the current research performed by
(name researcher)
Maxime Bos

I have agreed to take part in the study entitled
Participant behaviour in a modified Mod game
and I understand that my participation is entirely voluntary. I understand that my responses will be kept strictly confidential and anonymous. I have the option to withdraw from this study at any time without penalty, and I also have the right to request that my responses not be used.

The following points have been explained to me:

1. The goal of this study is
To learn what behaviour human participants show based on another player's actions
Participation in this study should help advance our understanding of
Human strategic behaviour

2. I shall be asked to
Play a 2-player Mod game on a computer for three blocks

3. The current study will last approximately **45-60** minutes. If desired, the researcher will explain to me in more detail what the research was about through email once all other participants have completed the experiment.

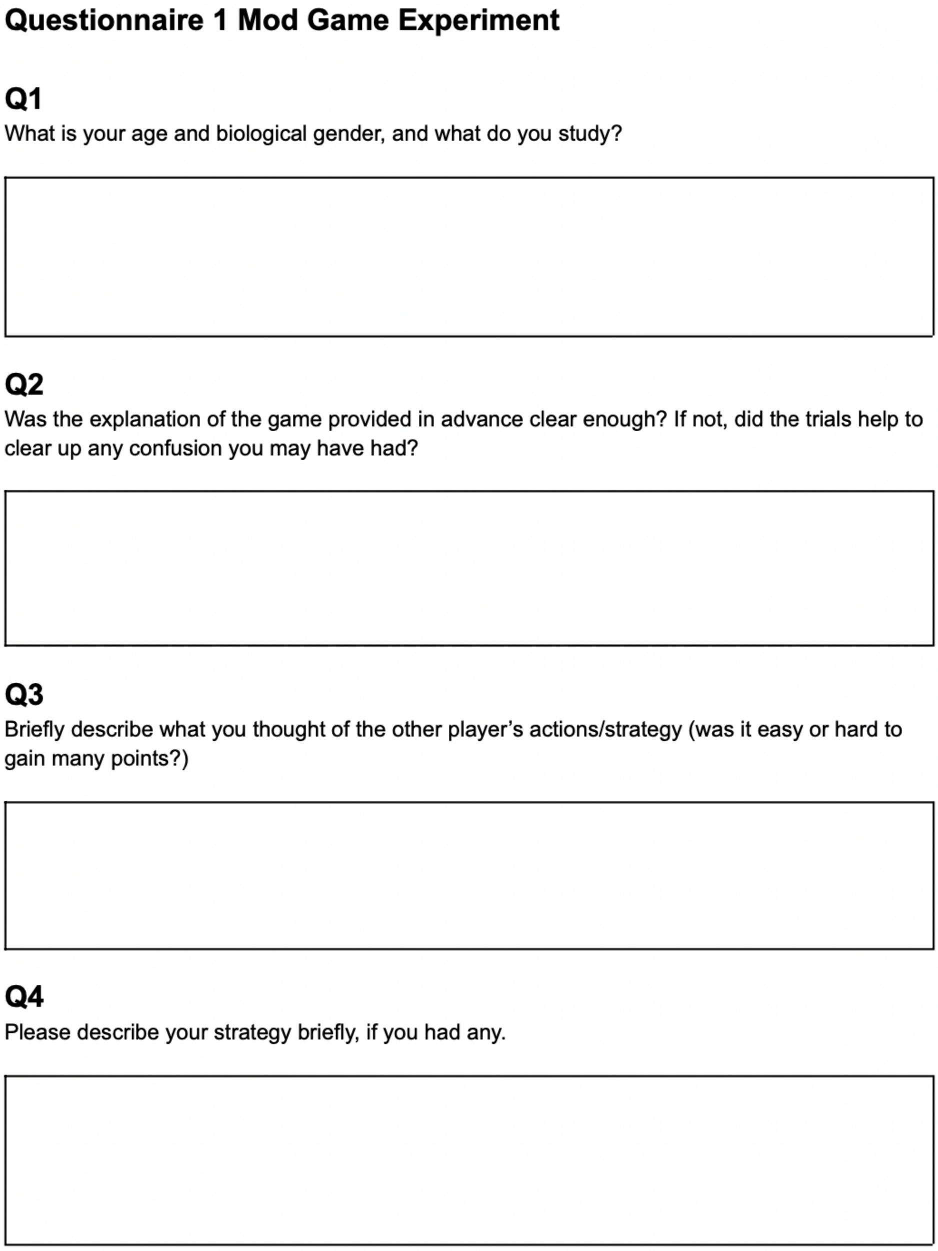
4. My responses will be treated confidentially, and my anonymity will be ensured. Hence, my responses cannot be identifiable and linked back to me as an individual.

5. The researcher will answer any questions I might have regarding this research (please leave your email so the researcher can reach you with any further information)

Date: Signature researcher:

Date: Signature participant:

Figure E.1: Informed consent form provided to participants before the start of the experiment



Questionnaire 1 Mod Game Experiment

Q1

What is your age and biological gender, and what do you study?

Q2

Was the explanation of the game provided in advance clear enough? If not, did the trials help to clear up any confusion you may have had?

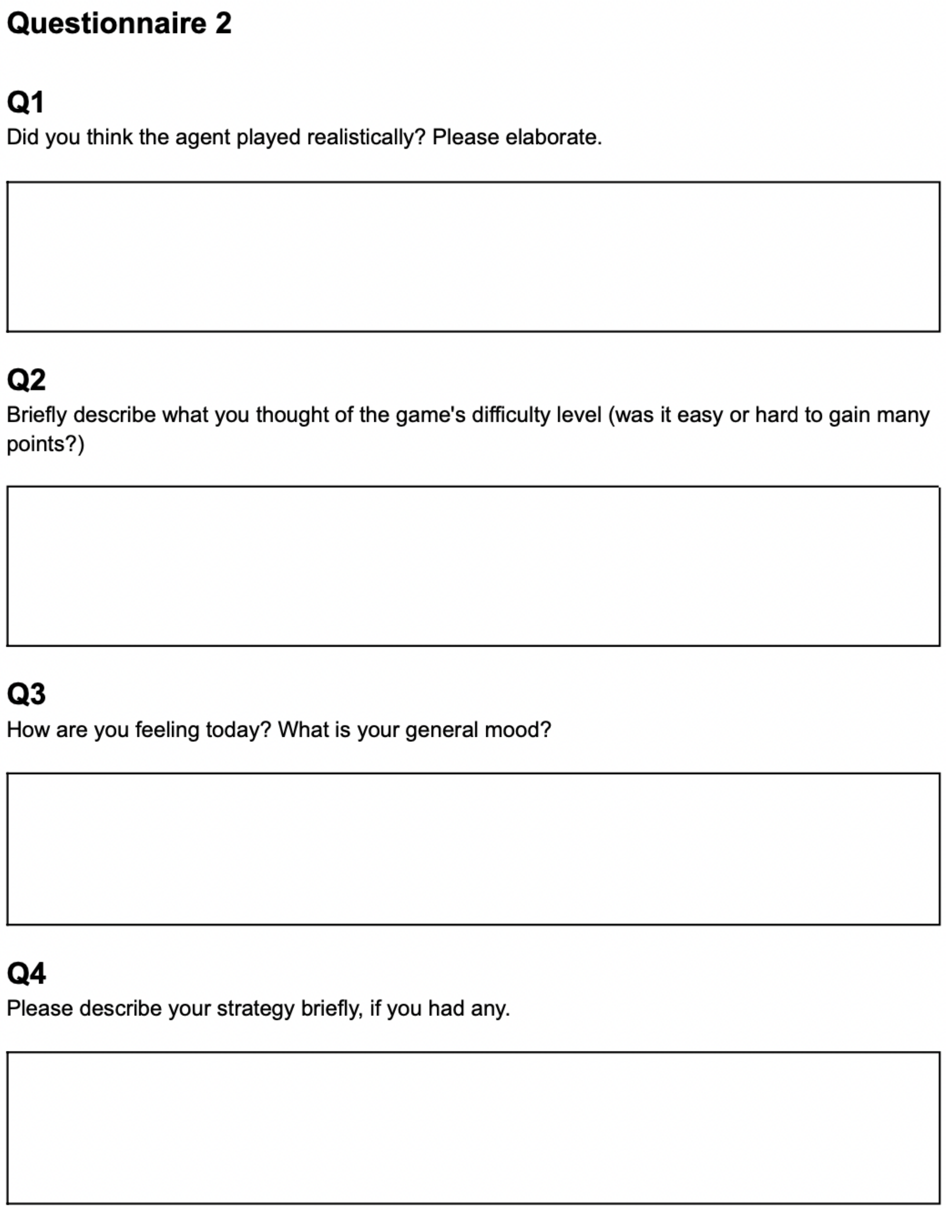
Q3

Briefly describe what you thought of the other player’s actions/strategy (was it easy or hard to gain many points?)

Q4

Please describe your strategy briefly, if you had any.

Figure E.2: Questionnaire 1 provided to participants to fill in during the experiment



Questionnaire 2

Q1

Did you think the agent played realistically? Please elaborate.

Q2

Briefly describe what you thought of the game's difficulty level (was it easy or hard to gain many points?)

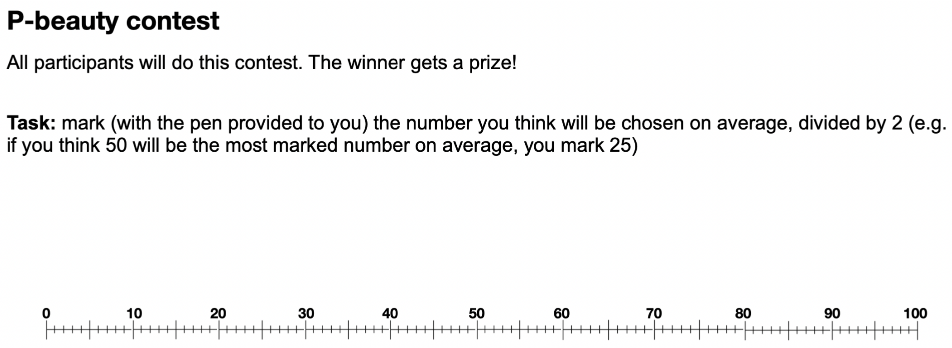
Q3

How are you feeling today? What is your general mood?

Q4

Please describe your strategy briefly, if you had any.

Figure E.3: Questionnaire 2 provided to participants to fill in during the experiment



P-beauty contest

All participants will do this contest. The winner gets a prize!

Task: mark (with the pen provided to you) the number you think will be chosen on average, divided by 2 (e.g. if you think 50 will be the most marked number on average, you mark 25)

0
10
20
30
40
50
60
70
80
90
100

Figure E.4: P-beauty contest provided to participants at the end of the experiment