university of groningen

faculty of mathematics
and natural sciences

Can a meta-cognitive model for a mixed-motive bargaining task outperform humans when contesting human players?

Bachelor Project Artificial Intelligence

June 2015

Student: J.D. Top

First supervisor: Dr. C.A. Stevens

Second supervisor: Prof.dr. N.A. Taatgen

Can a meta-cognitive model for a mixed-motive bargaining task outperform humans when contesting human players?

Jordi Top (s2402319)

July 10, 2015

Abstract

In this paper we investigate the question: “Can a meta-cognitive model for a mixed-motive bargaining task outperform humans when contesting human players?”. In our experiment two parties had to negotiate, with a computer model taking the role of one of the negotiators on half of our trials. Participants were asked to rate their counterpart’s agreeability, without knowing whether this was a human or the model. No significant difference in agreeability and absolute score was found, yet the model gained a significantly higher relative score. These findings suggest that even if it only helps to improve relative economical gains, teaching people the meta-cognitive strategy can help them become better negotiators, and will not impair their performance on other relevant performance measures.

Introduction

During a *negotiation*, two or more parties attempt to agree on a division of goods, earnings, costs, tasks or on a selling price for an item or service. Negotiations are an important part of our lives: we do not only use them when buying or selling goods, we also use them in a diverse set of other cases, such as dividing chores in a household, making a task division in group projects, splitting gas and electricity bills, deciding who can use the car at what time, and in many other things (Liebert et al. (1968), Galinsky and Mussweiler (2001)).

Previous work

In previous work, two negotiation strategies are usually distinguished: an aggressive, competitive or tough strategy, and a cooperative or soft strategy (e.g., Hüffmeier et al. (2014)). An *aggressive strategy* is used to maximize personal gains at the expense of the other negotiator(s), and is usually accompanied by a demanding first offer and few and small concessions (Yukl, 1974; Esser and Komorita, 1975). A *concession* occurs when a negotiator makes a new offer which decreases his gains or increases his losses, while, in return, increasing the gains or decreasing the losses of the other negotiator(s). A *cooperative strategy* aims to split the profits equally between all negotiators and tries to maximize the total profits across *all* negotiators. Someone using a cooperative strategy will, in general, make a less demanding initial offer and will make more frequent and larger concessions (Gray, 1977).

When trying to get the best economical gain from a negotiation, the aggressive strategy is usually recommended (Yukl, 1974; Huang et al., 2006), since small and few concessions and a demanding initial offer can lead to a more profitable final agreement. On the other hand, a cooperative strategy can lead to better socio-emotional outcomes, that is, a cooperative negotiator will be seen as more agreeable (Hüffmeier et al., 2014). This, in turn, can lead to more cooperation in the future.

This paper focuses on a third, more recent strategy: the *meta-cognitive strategy.* The meta-cognitive strategy employs *theory of mind*, that is, thinking about another person's beliefs and reasoning, to find out what the other negotiator's strategy is. The meta-cognitive negotiator then changes his own strategy based on the other negotiator's perceived strategy. It has been found that a meta-cognitive strategy outperforms a wide variety of other strategies in a three-agent cooperative game in Reitter et al. (2010). In Galinsky and Mussweiler (2001) it is found that taking the other negotiator's perspective can help counter biases in bargaining situations. Lastly, Zohar and Peled (2008) suggest that teaching of *meta-strategic reasoning*, reasoning about one's own strategy and adapting it where necessary, can improve experimentation abilities in elementary school pupils.

In this paper we investigate the effect of using a meta-cognitive strategy on the economical and socio-emotional outcomes of a negotiation. A purely cooperative negotiator cannot defend itself against an aggressive negotiator, whereas a meta-cognitive negotiator can resist exploitation by responding to toughness with toughness (Esser and Komorita, 1975). On the other hand, a purely aggressive negotiator cannot improve his socio-emotional outcomes and cannot improve his gains through cooperation when possible, and will force the other negotiator to also use an aggressive strategy, making the negotiation more difficult. Due to these shortcomings I suspect a meta-cognitive negotiator can get better economic and socio-emotional outcomes. If this is the case, it will be useful to teach people the meta-cognitive strategy.

To investigate these negotiation strategies, we use Kelley's *Game of Nines* (Kelley et al., 1967), a bargaining game where two negotiators have to divide a reward under incomplete information.

Model descriptions

Overview

Cognitive models capable of performing the Game of Nines task using each of the three strategies have been developed in the ACT-R cognitive architecture

(Anderson et al., 2004). They play by retrieving instances of an in-game situation from their memory. The meta-cognitive model uses these instance both to select its actions and to infers its opponent's strategy. Once the meta-cognitive model has inferred which strategy its opponent is using, it will employ the same strategy.

Detailed description

Each model plays by retrieving *chunks*, specifying a certain situation in the game, from their *declarative memory* to decide on their next action based on the other negotiator's last move and the distance between the model's current offer and his MNS. The chunk most similar to the current situation is retrieved. Since not all possible in-game situations are represented in chunks, partial matching is often required. With *partial matching*, a chunk is selected where one or more fields match the current situation, favoring chunks with more matching fields. Each model's initial set of chunks has been coded by hand, and is based on previous work regarding negotiations (Kelley et al. (1967), Liebert et al. (1968), Schoeninger and Wood (1969)). There are chunks corresponding to both strategies, as well as "neutral chunks", which fall between both strategies

The meta-cognitive model uses *instance-based learning*, learning by comparing new instances of a problem with instances previously encountered and stored in memory (see Aha et al. (1991)). It starts with cooperative, aggressive and neutral chunks, and has two "substrategies": cooperative and aggressive. It uses its chunks for two purposes: identifying its opponent's strategy, and selecting actions to perform. When identifying its opponent's strategy, it tries to match its opponent's actions with their most similar chunks. If this is an aggressive chunk, it can infer that its opponent is using and aggressive strategy, and vice versa for cooperative chunks. Neutral chunks ensure ambiguous actions aren't classified as aggressive or cooperative. When the model recognizes its opponent's strategy as aggressive or cooperative, it will switch to its corresponding substrategy. Neutral actions are ignored. Like the two other models, the meta-cognitive model matches the current situation with chunks in its memory to select an action. However, selecting chunks highly depends on the substrategy the model is currently using: chunks corresponding to its current strategy have a high probability of being selected, neutral chunks have a low probability and chunks corresponding to the substrategy it is currently *not* using have a very slim chance of being selected. Using this structure, the meta-cognitive model *reciprocates*, that is, matches its opponent's strategy.

The cooperative chunks specify less demanding initial offers and lower *lowest acceptable gains*, the smallest gains it will still agree to, whereas the aggressive chunks specify more demanding initial offers and higher lowest acceptable gains. Neutral chunks have intermediate initial offers as well as intermediate lowest acceptable gains. For a more complete description, see Stevens (2015).

Previous experiment

In a previous experiment (Stevens, 2015), the three models played against two agents. The agents used formulae to calculate their next move. The *fair* agent tried to equally split the profits between himself and the other negotiator, whereas the *unfair* agent tried to maximize his profits at the expense of the other negotiator. It was found that the aggressive and cooperative model performed better against the unfair and fair agent, respectively. However, the meta-cognitive model performed equal to or better than the other two models against either agent. Moreover, when compared with human performance against both agents, the meta-cognitive model performed as well as the top 25% of the participants. This substantiates our suspicions that a meta-cognitive negotiation strategy can yield a better economic outcome than the aggressive or cooperative strategy alone, and provides some evidence that teaching people the meta-cognitive strategy will make them better negotiators.

Research question

In this paper we build on these previous findings by comparing the meta-cognitive model with humans when playing against (other) human negotiators. We aim to answer the question "Can a meta-cognitive model for a mixed-motive bargaining task outperform humans when contesting human players?", with "performance" referring to both profits and socio-emotional outcomes. Since we also wish to know how well the meta-cognitive model represents a human negotiator we'll use a set-up similar to a Turing test (Turing, 1950). This will also help us in measuring socio-emotional outcomes, as "agreeability", when describing another negotiator, might have a different meaning when the other negotiator is perceived as a computer instead of a human.

Method

Overview

In each experiment, two players played the Game of Nines against each other over fourteen rounds. Our primary manipulation was whether player 2 played against a human partner or a confederate operating the meta-cognitive model. Player 2 was never informed whether he was playing against a human or the model.

The game

The game which was used is Kelley's *Game of Nines* (Kelley et al., 1967). In the Game of Nines, two negotiators had to agree on a division of nine points. However, both participants also received a *Minimum Necessary Share* (MNS) which was subtracted from their part of the agreed division. Both negotiators only knew their own MNS, and *were not allowed* to reveal it to the other negotiator. If a negotiator agreed on receiving a number of points *under* his MNS, he would receive a negative number of points which was subtracted from his points acquired over multiple bargaining rounds with the other negotiator. Single points were not divisible: both negotiators had to agree on a whole number of points on each round. Points could also not be "left on the table": all nine points had to be divided between the negotiators. Negotiators could quit during a negotiation: if one of the negotiators quit during a round, both received zero points for this round, regardless of their MNS. To limit the total duration of a trial, each round could only take three minutes. If these three minutes were exceeded, both participants received zero points, regardless of their MNS, for the current round.

Introduction

Each trial was performed with two participants, a *dyad*. Before a trial started, the game was explained to the participants. The participants were asked if they had any more questions about the rules, and if they understood them. The participants were told one of them would be taking the role of the *confederate*, who would either play by himself or would control a model. The other player, to be referred to as the *player*, always played by himself. To eliminate any effects the messages might have had on agreeability ratings, both participants had to use a predefined set of messages to communicate, one for each action (these are "Deal.", "I quit.", "Final offer." and the numbers 1 through 9). They received a sheet with these messages for quickly looking them up. Since the model can only play in a turn-based Game of Nines, the players had to take turns performing actions. First the participants played three introductory rounds to ensure they understood the rules. The points gained during these rounds were discarded, and the model, if used, was reset before the actual rounds started. To prevent priming effects (as found by Burnham et al. (2000)) the term "counterpart" was always used when describing the other negotiator.
Before any rounds were played (this includes introductory rounds), both negotiators were separated so they could not see or hear each other. If the confederate operated the model, he *had to* use the same moves as the model.

Experimental set-up

Negotiation was performed using an open source instant messaging client called LAN Messenger. During the experiment, three channels were used: one for the player and the experimenter, which the experimenter used to send the player his MNS and score, one for the confederate and the experimenter which was used in a similar manner, and one shared between all three parties, which was used for negotiation between the player and confederate, and for announcing who would make the first offer, which round is being played and the division of points at the end of each round.

Experimental conditions

Each set of two participants played fourteen rounds, using the following set of tuples:
(1,1) (2,2) (3,3) (4,4) (1,3) (3,1) (1,5) (5,1) (3,4) (4,3) (2,6) (6,2) (4,5) (5,4)
To ensure neither party gained a "low man's advantage" (Kelley et al., 1967) during a block, an advantage over the other player because your MNS values are lower than his, we always used both the original and the mirrored MNS tuple for each tuple with unequal MNS values. To prevent order effects the tuple order was randomized for each set of participants. Participants took turns in making the first offer.
There were two conditions: either the confederate played by himself or he operated a model. The first will be referred to as the "human vs. human condition", abbreviated "hvh" whereas the latter is the "human vs. model condition", or "hvm".

Participants

Thirty-eight participants were recruited from a Facebook group for people who are interested in participating in paid experiments in Groningen. Twenty of these played in the human vs. human condition and the other eighten played in the human vs. model condition. In the human vs. model condition, four confederates were used, of which two confederates were used for most trials, without the other participant knowing their counter-player had participated before. There were ten dyads, and thus ten trials, for the human vs. human condition, and fourteen trials for the human vs. model condition. The participants were given ten euros for participating, the two returning confederates for the human vs. model trials were given ten euros for each trial they participated in.

Evaluation

After each trial, the non-confederate player was given a questionnaire asking to rate their counterpart's agreeability and how much they suspected they were playing against a human player on a scale from 1 to 10. Three different questions were used to rate agreeability, and one question was used to rate "humanity". The questions were the following:

- Based on the actions of the other negotiator, how "agreeable" was this negotiator on a scale from 1 to 10? 1 means they weren't agreeable at all, 10 means they were incredibly agreeable.
- How much did you enjoy playing against the other negotiator on a scale from 1 to 10? 1 means you didn't enjoy it at all, 10 means you enjoyed it a lot.
- Did you like the other player's strategy on a scale from 1 to 10? 1 means you didn't like it at all, 10 means you liked it a lot.

- On a scale from 1 to 10, how much do you think you were playing against a human? 1 means you're absolutely certain it was a computer model, 10 means you're absolutely certain it was the other participant.

Data to be collected

Several points of data were collected: first of all, ratings of agreeability and humanity were explicitly requested after each trial. Secondly, the total number of points earned by the player and model, the number of rounds quit by each player and the number of final offers made by each player were tracked throughout the experiment. For each human vs. model trial, a factor was calculated specifying the number of turns in which the model was using its cooperative substrategy, divided by the total number of turns. This factor specifies how cooperative the model was during a trial: if it was cooperative on each turn it would have a value of one, if it was aggressive on each turn it would have a value of zero.

Results

In total, twenty-four pairs of subjects participated in the experiment, ten in the human vs. human condition and fourteen in the human vs. model condition. In three of the fourteen human vs. model trials, the model's operator made an error. These trials have been excluded from our data analysis, leaving eleven human vs. model trials for analysis.
In each trial, the following data was collected: the condition, each player's final score, each player's number of final offers, the number of times each player has quit, three questionnaire ratings on agreeability and one questionnaire rating on humanity. For human vs. model trials, the model's cooperativeness was calculated, as discussed more thoroughly in the previous section.
The distribution of players across conditions can be found in Table 1 on page 8. It can be seen that player 1 in the human vs. model condition was always the model, which was operated by a confederate. All other players were actual participants. In our analysis we use the following terminology for several subgroups of participants: "the model" is player 1 in the hvm condition. "all humans" are all cells *except* the player 1, hvm cell. "The model's counterparts" are all humans who played against the model, so the hvm, player 2 cell. "hvh players" are all players who played in the hvh condition, so the union of the player 1, hvh cell and the player 2, hvh cell.

	player 1	player 2
human vs. human	human	human
human vs. model	model	human

Table 1: Conditions and participants

Exploratory data analysis

For total scores, the means, minima and maxima for each (sub)group of players can be seen in Table 2 on page 9. The standard deviations are displayed alongside the means between brackets. All values have been rounded to two decimals. Although the model's mean score is higher than its counterparts and all humans, it stays under the mean score of all hvh players.

The total number of points which could be obtained over all fourteen trials is equal to 9 × 14, minus the sum of all MNS values, 88, so the total number of points available is 38. If two players were perfectly cooperative, they could obtain 19 points each. In certain rounds one player had to accept gaining zero points, whereas the other would gain only one point. This, however, is very unlikely as players often reject any request in which they do not gain at least one point. If someone had acquired more than 19 points, it is likely this player has taken advantage of his counterpart.

	all	model	model counterparts	all humans	hvh players
mean	13.12 (4.14)	13.91 (3.62)	9.73 (2.95)	12.84 (4.33)	14.55 (4.05)
minimum	3	7	3	3	6
maximum	21	21	13	21	21

Table 2: Means, maxima and minima

To get more insight into the data and the average behaviour of players we also looked at mean quitting and final offers, as seen in Table 3 on page 9, again all rounded to two decimals, with standard deviations between brackets. Over all trials, a player quit 3.71 times on average, so 2 × 3.71 = 7.42 rounds were quit in total, on average. In the rounds with MNS tuples (4,5) and (5,4), no points could be obtained, so quitting was to be expected. In the rounds with MNS tuples (4,4), (6,2) and (2,6) one player had to agree to obtaining zero points, so quitting also occured very often (although there have been trials in which participants reached an agreement in these rounds). In all (sub)groups of participants, the mean number of final requests was higher than the mean number of quits. In very few rounds participants quit without a final request.

	all	model	counterparts	humans	hvh players
mean quits	3.71 (1.74)	4.27 (1.95)	4.45 (1.98)	3.52 (1.65)	3.00 (1.21)
mean final requests	4.93 (2.09)	5.73 (2.61)	5.09 (2.34)	4.65 (1.84)	4.40 (1.50)

Table 3: Mean quitting and final offers

The three questions on agreeability or denoted are "agr1", "agr2" and "agr3" respectively. The humanity score is denoted as "hum". Player 2 filled in the questionnaire concerning player 1, so there are only agreeability and humanity ratings concerning player 1. Mean questionnaire ratings can be found in Table 4 on page 10. Again, all values have been rounded to two decimals and standard deviations are displayed between brackets. It can be seen that on average,

	all	model	hvh player 1
mean agr1	4.90 (1.73)	4.36 (1.91)	5.50 (1.35)
mean agr2	6.67 (1.71)	6.45 (2.30)	6.90 (0.74)
mean agr3	4.57 (1.96)	4.45 (2.46)	4.70 (1.34)
mean hum	5.71 (2.45)	5.82 (2.82)	5.60 (2.12)

Table 4: Mean questionnaire results

human players have been rated as more agreeable across all three questions on agreeability. However, the mean humanity rating is higher for the model.

Lastly, the mean cooperativity factor of the model was approximately 0.33, with a standard deviation of approximately 0.29, suggesting the model used its aggressive substrategy in about two-thirds of its turns throughout the entire experiment. This factor ranged between 0.08 and 0.93: against some players it played almost exclusively aggressively, whereas against others it played very cooperatively.

Statistical analysis

Scores

In our statistical analysis we compared the model's mean score with both the human counterparts and all hvh players. To perform a t-test, data must be drawn from a normal distribution. To test whether the data is drawn from a normal distribution, we used a Shapiro-Wilk test of normality over all total scores. In this test, and in all further tests, we used a significance threshold of $\alpha = 0.05$. According to a Shapiro-Wilk test's null-hypothesis, the data is normally distributed. We obtained a non-significant p-value with $W = 0.97$ and $p > 0.05$. We can not reject the null-hypothesis, so we assume the data is drawn from a normal distribution.

First we performed a comparison of means between the model's scores (μ_m) and the scores of all players in the hvh condition (μ_h) with $H_0 : \mu_h = \mu_m$and $H_a : \mu_h \neq \mu_m$, using a Welch two-sample t-test. The model's scores did not differ significantly from the score in the hvh condition, with $t(22.80) = -0.45$.

Secondly we compared the mean of the model's scores with the means of the model's counterpart's scores. Whereas the previous test can be seen as a comparison of absolute score, this test looks at relative score. The mean of the model's counterpart's scores is denoted as μ_c. Our null-hypothesis was $H_0 : \mu_m = \mu_c$, our alternative hypothesis was two-sided, $H_a : \mu_m \neq \mu_c$, once again we used a Welch two-sample t-test. The model's scores differed significantly from the model's counterpart's scores, with $t(19.19) = 2.98$ and $p = 0.007693$. To further investigate this difference, we performed another Welch two-sample t-test, this time using $H_a : \mu_m > \mu_c$. The model's score is significantly greater than the model's counterpart's score, with $t(19.19) = 2.98$ and $p = 0.003847$.

Agreeability

At the end of each trial, player 2 was asked to fill out a questionnaire concerning player 1, before revealing to player 2 whether he was playing against another player or the model. Before comparing agreeability scores, we ensured the model could not be discerned from human players, and therefore the agreeability rating was not influenced by this knowledge. A Shapiro-Wilk test of normality on all humanity ratings resulted in $W = 0.95$ with $p > 0.05$, so we can not accept the test's alternative hypothesis that the data is not drawn from a normal distribution, and assume it is. We used a Welch two-sample t-test on the mean humanity score for the model (μ_m) and the human players (μ_h). Our null-hypothesis was $H_0 : \mu_h = \mu_m$, our alternative hypothesis was $H_a : \mu_h \neq \mu_m$. The model's humanity rating did not differ significantly from the human humanity rating, with t(18.39)= -0.20146.

First, we tested the correlations between all three agreeability ratings, to see if they could be combined. We performed a Pearson's product-moment correlation test on each combination of two agreeability ratings, and used $H_a : R > 0$ as alternative hypotheses. The results of this test can be found in Table 5 on page 11, with correlation coefficients R rounded to two decimals. According to Table 5, each combination of agreeability ratings is significantly positively correlated, with p-values of $p < 0.05$ on each test. Since all three agreeability ratings are positively correlated, we computed the mean agreeability for each trial and used these in our statistical analysis.

	agr1 and agr2	agr1 and agr3	agr2 and agr3
R	0.71	0.65	0.73
t-value	4.45	3.73	4.63
degrees of freedom	19	19	19
p	0.0001362	0.0007144	9.057×10^{-5}

Table 5: Correlation test results for agreeability ratings

We once again used a Shapiro-Wilk test of normality to test if the mean agreeability ratings are drawn from a normal distribution. We found $W = 0.97$ with $p > 0.05$, so we could not reject the null-hypothesis that the data is drawn from a normal distribution, and proceeded with a comparison of means between the model's and human mean agreeability. We performed a Welch two-sample t-test, using $H_a : \mu_h \neq \mu_m$. There was no significant difference between mean agreeability for the model and the human player with t(14.53) = 0.89.

Discussion

The results of our experiment show that the meta-cognitive model does not perform significantly worse than human players on any of our relevant metrics. Our experiment adheres to the previous experiment (Stevens (2015)) as discussed in the introduction, as it either performs equal to or better than the other players.

As mentioned in the results section, we tested both *absolute* and *relative* differences in score. A relative difference indicates that the model obtains less or more points than its counterpart, regardless of their total amount of points: it shows which player "beat" the other player. An absolute difference indicates who obtains the highest total gain when pitted against others instead of each other. No significant difference of means was found in absolute scores of the model and the human players, indicating the model can gain as much points as others in negotiations.

We *did* observe a significant difference of means in relative scores of the model and human players. More specifically, the model's mean score was significantly higher than their counterparts' mean score. This indicates the model is adept at "beating" its opponents. This adheres to the findings in (Stevens, 2015), where the meta-cognitive model fits the data of the top quartile of human participants: the model is better than the average participant.

In an actual negotiation setting, absolute gain can be more important than relative gain. For example, most people would prefer a deal where they gain twenty euros and the other gains twenty-five euros over a deal where they gain ten euros and the other gains five euros. In the latter, they have beaten the other negotiator, but this leaves them with less total gain. Overall we can say the meta-cognitive model's economic outcome is equal to or better than the economic outcome of our average participant, as we aspired. From this we might deduce that the meta-cognitive strategy can provide better economic outcomes, or at least better relative economic outcomes.

Our results also indicate that the meta-cognitive model, if disguised properly, cannot be significantly distinguished from human players, which may be useful for future experiments concerning socio-emotional performance of this model.

We did not observe a significant difference between the model and human players concerning mean agreeability. We could infer that the model is not *less* agreeable than human players, so the meta-cognitive strategy's socio-emotional outcomes are not worse than those of human players. In future research, the model's socio-emotional gains could be compared to those of a purely cooperative or aggressive model, which would provide proof that the meta-cognitive strategy, even if it sometimes uses aggressive actions, performs equal to or better than the cooperative strategy concerning socio-emotional outcomes.

We set out to provide evidence that teaching people the meta-cognitive strategy can help them become better negotiators. This paper supports this statement: although the meta-cognitive strategy may not have obtained a better absolute economic outcome or a better socio-emotional outcome, it did achieve a better relative economic outcome. On none of these metrics the meta-cognitive model did worse than the human negotiators, so even if it only improves their relative economic outcome, teaching people the meta-cognitive strategy will still benefit them.

References

Aha, D. W., Kibler, D., and Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning*, 6:37–66.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., and Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, 111(4):1036–1060.

Burnham, T., McCabe, K., and Smith, V. L. (2000). Friend-or-foe intentionality priming in an extensive form trust game. *Journal of Economic Behavour & Organization*, 43:57–73.

Esser, J. K. and Komorita, S. S. (1975). Reciprocity and concession making in bargaining. *Journal of Personality and Social Psychology*, 31(5):864–872.

Galinsky, A. D. and Mussweiler, T. (2001). First offers as anchors: The roles of perspective-taking and negotiator focus. *Journal of Personality and Social Psychology*, 81(4):657–669.

Gray, S. H. (1977). Model predictability in bargaining. *Journal of Psychology*, 97(2):171–178.

Huang, S., Lin, F., and Yuan, Y. (2006). Understanding agent-based on-line persuasion and bargaining strategies: An empirical study. *International Journal of Electronic Commerce*, 11(1):85–115.

Hüffmeier, J., Freund, P. A., Zerres, A., Backhaus, K., and Hertel, G. (2014). Being tough or being nice? a meta-analysis on the impact of hard- and softline strategies in distributive negotiations. *Journal of Management*, 40(3):866–892.

Kelley, H. H., Beckman, L. L., and Fischer, C. S. (1967). Negotiating the division of a reward under incomplete information. *Journal of Experimental Social Psychology*, 3:361–398.

Liebert, R. M., Smith, W. P., Hill, J. H., and Keiffer, M. (1968). The effects of information magnitude of initial offer on interpersonal negotiation. *Journal of Experimental Social Psychology*, 4:431–441.

Reitter, D., Juvina, I., Stocco, A., and Lebiere, C. (2010). Resistance is futile: Winning lemonade market share through metacognitive reasoning in a three-agent cooperative game. In *Proceedings of the 19th Conference on Behavioral Representation in Modeling and Simulation (BRIMS)*, Charleston, S.C.

Schoeninger, D. W. and Wood, W. D. (1969). Comparison of married and ad hoc mixed-sex dyads negotiating the division of a reward. *Journal of Experimental Social Psychology*, 5:483–499.

Stevens, C. (2015). Cognitive model of the game of nines. *Paper in preparation.*

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59:433–460.

Yukl, G. (1974). Effects of the opponent's initial offer, concession magnitude and concession frequency on bargaining behavior. *Journal of Personality and Social Psychology*, 30(3):323–335.

Zohar, A. and Peled, B. (2008). The effects of explicit teaching of metastrategic knowledge on low- and high-achieving students. *Learning and Instruction*, 18:337–353.

Appendices

Instructions and questionnaire

The instruction sheet, which also contains the questionnaire, can be seen on Figures 1 through 6 on pages 1, 2, 3, 4, 5 and 6.

Data in .csv-format

In this section all data is presented in .csv-format.

```
”p1total”,”p2total”,”p1quits”,”p2quits”,”coop”,”p1finals”,”p2finals”,”game”,”mnum”,”hnum”,”agr1”,”agr2”,”agr3”,”hum”,”void”
12,10,7,3,0.928571428571429,1,7,2,1,0,3,4,4,3,1
28,3,0,6,0.916666666666667,0,0,2,2,0,6,7,5,2,1
14,10,1,7,0.117117117117117,11,2,2,3,0,3,5,5,4,0
14,9,5,4,0.19047619047619,6,5,2,4,0,2,3,2,4,0
11,7,7,3,0.0823529411764706,4,9,2,5,0,5,9,4,1,0
14,13,6,2,0.584269662921348,4,6,2,6,0,4,5,4,10,0
16,12,5,3,0.181818181818182,4,8,2,7,0,7,9,6,10,0
13,12,3,6,0.464788732394366,7,2,2,8,0,2,4,1,5,0
7,10,5,6,0.154929577464789,7,7,2,9,0,3,6,5,4,0
11,9,2,7,0.0909090909090909,8,4,2,10,0,6,8,4,7,0
21,3,2,6,0.115789473684211,7,3,2,11,0,6,10,10,7,0
15,13,6,2,0.703703703703704,2,6,2,12,0,3,5,2,4,0
17,9,5,3,0.933333333333333,3,4,2,13,0,7,7,6,8,0
12,11,5,3,0.860759493670886,1,4,2,14,0,5,7,5,4,1
16,20,2,2,-1,2,3,1,0,1,7,7,5,6,0
6,17,2,3,-1,5,5,1,0,2,6,8,5,7,0
7,15,3,5,-1,7,3,1,0,3,6,7,5,3,0
16,12,2,4,-1,4,4,1,0,4,5,7,4,5,0
14,16,5,2,-1,3,6,1,0,5,4,6,2,8,0
13,21,3,2,-1,4,6,1,0,6,7,8,7,9,0
```

```
14,8,6,3,−1,4,5,1,0,7,3,7,4,4,0
16,16,3,2,−1,6,5,1,0,8,7,6,5,3,0
16,16,2,4,−1,5,6,1,0,9,5,6,4,4,0
20,12,2,3,−1,4,1,1,0,10,5,7,6,7,0
```

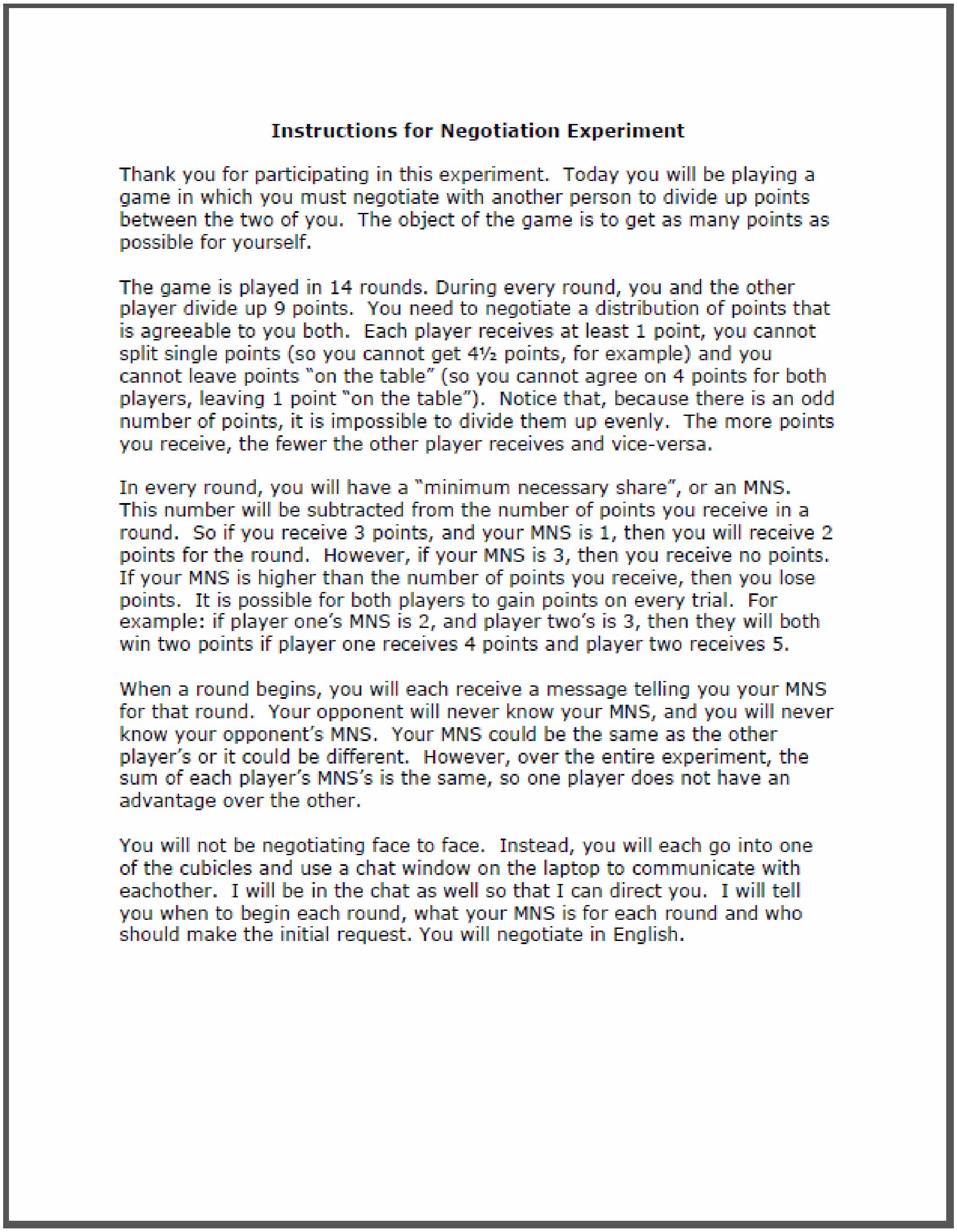
R code

The R code used in our statistical analysis is as follows:

```
#Reading the data from newlogs.csv
readfile = read.csv(file=”newlogs.csv”,head=TRUE, sep=”,”)

#Removing void trials
newlogs = readfile[which(newlogs$void == ”0”),]

#Adding mean agreeability
newlogs$agrmean = (newlogs$agr1 + newlogs$agr2 + newlogs$agr3)/3

#Tests and values, uncomment to perform a test.

#Mean, standard deviation, minimum and maximum points of subgroups
#all
#mean(c(newlogs$p1total,newlogs$p2total))
#sd(c(newlogs$p1total,newlogs$p2total))
#min(c(newlogs$p1total,newlogs$p2total))
#max(c(newlogs$p1total,newlogs$p2total))
#model
#mean(newlogs$p1total[which(newlogs$game==2)])
#sd(newlogs$p1total[which(newlogs$game==2)])
#min(newlogs$p1total[which(newlogs$game==2)])
#max(newlogs$p1total[which(newlogs$game==2)])
#model counterparts
#mean(newlogs$p2total[which(newlogs$game==2)])
#sd(newlogs$p2total[which(newlogs$game==2)])
#min(newlogs$p2total[which(newlogs$game==2)])
#max(newlogs$p2total[which(newlogs$game==2)])
#all humans
#mean(c(newlogs$p2total,newlogs$p1total[which(newlogs$game==1)]))
#sd(c(newlogs$p2total,newlogs$p1total[which(newlogs$game==1)]))
#min(c(newlogs$p2total,newlogs$p1total[which(newlogs$game==1)]))
```

```
#max(c(newlogs$p2total, newlogs$p1total[which(newlogs$game==1)]))
#hvh players
#mean(c(newlogs$p1total[which(newlogs$game==1)], newlogs$p2total[which(newlogs$game==1)]))
#sd(c(newlogs$p1total[which(newlogs$game==1)], newlogs$p2total[which(newlogs$game==1)]))
#min(c(newlogs$p1total[which(newlogs$game==1)], newlogs$p2total[which(newlogs$game==1)]))
#max(c(newlogs$p1total[which(newlogs$game==1)], newlogs$p2total[which(newlogs$game==1)]))

#mean quits and final requests for each subgroup, and standard deviations
#all
#mean(c(newlogs$p1quits, newlogs$p2quits))
#sd(c(newlogs$p1quits, newlogs$p2quits))
#mean(c(newlogs$p1finals, newlogs$p2finals))
#sd(c(newlogs$p1finals, newlogs$p2finals))
#model
#mean(newlogs$p1quits[which(newlogs$game==2)])
#sd(newlogs$p1quits[which(newlogs$game==2)])
#mean(newlogs$p1finals[which(newlogs$game==2)])
#sd(newlogs$p1finals[which(newlogs$game==2)])
#model counterparts
#mean(newlogs$p2quits[which(newlogs$game==2)])
#sd(newlogs$p2quits[which(newlogs$game==2)])
#mean(newlogs$p2finals[which(newlogs$game==2)])
#sd(newlogs$p2finals[which(newlogs$game==2)])
#all humans
#mean(c(newlogs$p2quits, newlogs$p1quits[which(newlogs$game==1)]))
#sd(c(newlogs$p2quits, newlogs$p1quits[which(newlogs$game==1)]))
#mean(c(newlogs$p2finals, newlogs$p1finals[which(newlogs$game==1)]))
#sd(c(newlogs$p2finals, newlogs$p1finals[which(newlogs$game==1)]))
#hvh players
#mean(c(newlogs$p1quits[which(newlogs$game==1)], newlogs$p2quits[which(newlogs$game==1)]))
#sd(c(newlogs$p1quits[which(newlogs$game==1)], newlogs$p2quits[which(newlogs$game==1)]))
#mean(c(newlogs$p1finals[which(newlogs$game==1)], newlogs$p2finals[which(newlogs$game==1)]))
#sd(c(newlogs$p1finals[which(newlogs$game==1)], newlogs$
    p2finals[which(newlogs$game==1)]))

#mean agr1, agr2, agr3 and hum for each hvm subgroup
#all
#mean(newlogs$agr1)
#mean(newlogs$agr2)
#mean(newlogs$agr3)
#mean(newlogs$hum)
#model
#mean(newlogs$agr1[which(newlogs$game==2)])
#mean(newlogs$agr2[which(newlogs$game==2)])
#mean(newlogs$agr3[which(newlogs$game==2)])
#mean(newlogs$hum[which(newlogs$game==2)])
#hvh player 1
#mean(newlogs$agr1[which(newlogs$game==1)])
#mean(newlogs$agr2[which(newlogs$game==1)])
#mean(newlogs$agr3[which(newlogs$game==1)])
#mean(newlogs$hum[which(newlogs$game==1)])

#Standard deviations for agr1, agr2, agr3 and hum for each hvm subgroup
#all
#sd(newlogs$agr1)
#sd(newlogs$agr2)
#sd(newlogs$agr3)
#sd(newlogs$hum)
#model
#sd(newlogs$agr1[which(newlogs$game==2)])
#sd(newlogs$agr2[which(newlogs$game==2)])
#sd(newlogs$agr3[which(newlogs$game==2)])
#sd(newlogs$hum[which(newlogs$game==2)])
#hvh player 1
#sd(newlogs$agr1[which(newlogs$game==1)])
#sd(newlogs$agr2[which(newlogs$game==1)])
#sd(newlogs$agr3[which(newlogs$game==1)])
#sd(newlogs$hum[which(newlogs$game==1)])

#mean and range of cooperativity
#mean(newlogs$coop[which(newlogs$game==2)])
#sd(newlogs$coop[which(newlogs$game==2)])
#min(newlogs$coop[which(newlogs$game==2)])
#max(newlogs$coop[which(newlogs$game==2)])

#Test for normality
#shapiro.test(c(newlogs$p1total,newlogs$p2total))

```

```
#Compare model to all hvh players
#t.test(newlogs$p1total[which(newlogs$game==2)],c(newlogs$p1total[which(newlogs$game==1)],newlogs$p2total[which(newlogs$game==1)]),alt=”two.sided”)

#Compare model to counterpart
#t.test(newlogs$p1total[which(newlogs$game==2)],newlogs$p2total[which(newlogs$game==2)],alt=”two.sided”)
#t.test(newlogs$p1total[which(newlogs$game==2)],newlogs$p2total[which(newlogs$game==2)],alt=”greater”)

#Normality of humanity
#shapiro.test(newlogs$hum)

#humanity means are equal
#t.test(newlogs$hum[which(newlogs$game==1)],newlogs$hum[which(newlogs$game==2)])

#Normality of agreeability
#shapiro.test(c(newlogs$agr1,newlogs$agr2,newlogs$agr3))

#Test correlation
#cor.test(newlogs$agr1,newlogs$agr2,alt=”greater”)
#cor.test(newlogs$agr1,newlogs$agr3,alt=”greater”)
#cor.test(newlogs$agr2,newlogs$agr3,alt=”greater”)

#Normality of mean agreeability
#shapiro.test(newlogs$agrmean)

#Compare model and human mean agreeability
#t.test(newlogs$agrmean[which(newlogs$game==1)],newlogs$agrmean[which(newlogs$game==2)])
```



Instructions for Negotiation Experiment

Thank you for participating in this experiment. Today you will be playing a game in which you must negotiate with another person to divide up points between the two of you. The object of the game is to get as many points as possible for yourself.

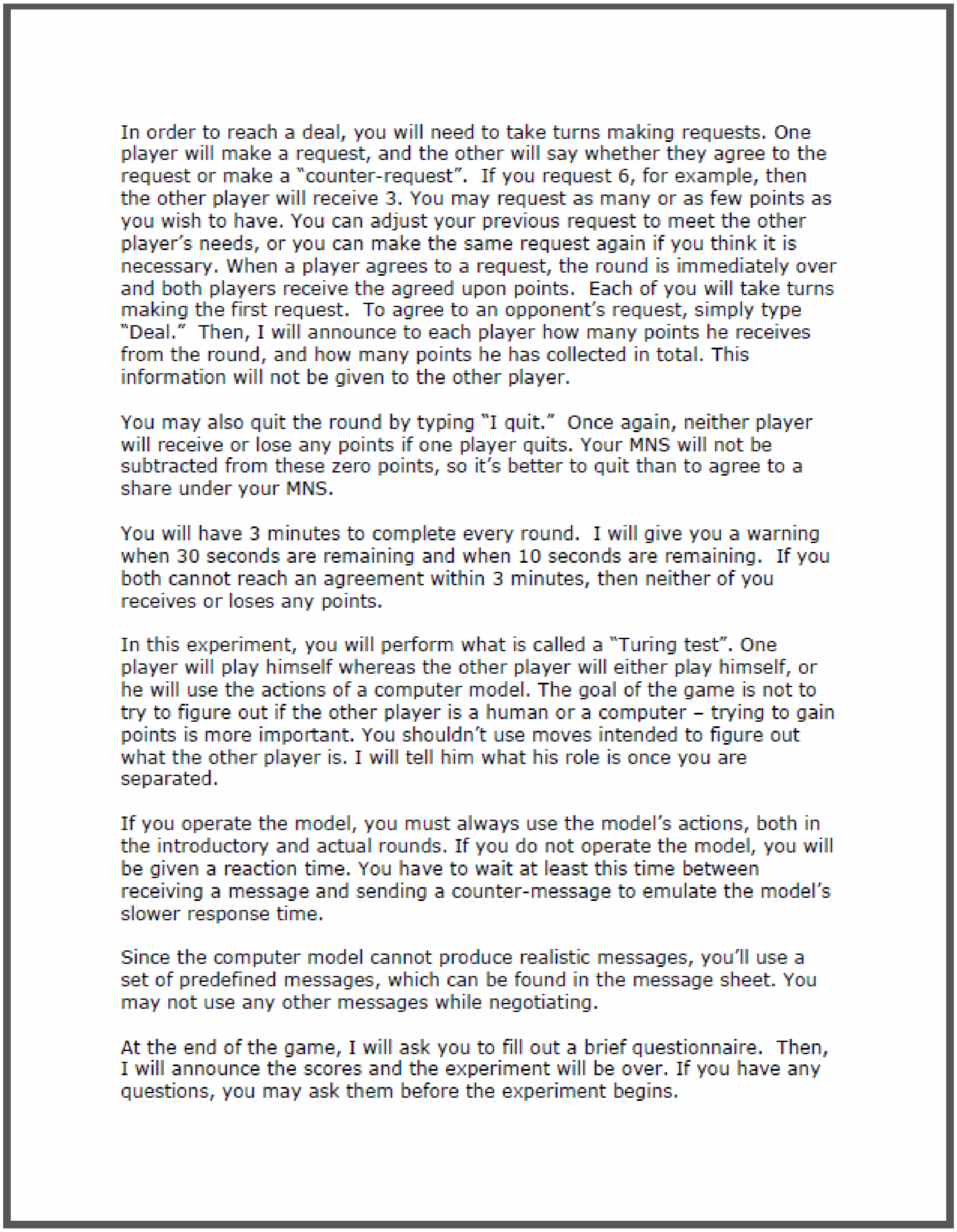
The game is played in 14 rounds. During every round, you and the other player divide up 9 points. You need to negotiate a distribution of points that is agreeable to you both. Each player receives at least 1 point, you cannot split single points (so you cannot get 4½ points, for example) and you cannot leave points “on the table” (so you cannot agree on 4 points for both players, leaving 1 point “on the table”). Notice that, because there is an odd number of points, it is impossible to divide them up evenly. The more points you receive, the fewer the other player receives and vice-versa.

In every round, you will have a “minimum necessary share”, or an MNS. This number will be subtracted from the number of points you receive in a round. So if you receive 3 points, and your MNS is 1, then you will receive 2 points for the round. However, if your MNS is 3, then you receive no points. If your MNS is higher than the number of points you receive, then you lose points. It is possible for both players to gain points on every trial. For example: if player one’s MNS is 2, and player two’s is 3, then they will both win two points if player one receives 4 points and player two receives 5.

When a round begins, you will each receive a message telling you your MNS for that round. Your opponent will never know your MNS, and you will never know your opponent’s MNS. Your MNS could be the same as the other player’s or it could be different. However, over the entire experiment, the sum of each player’s MNS’s is the same, so one player does not have an advantage over the other.

You will not be negotiating face to face. Instead, you will each go into one of the cubicles and use a chat window on the laptop to communicate with eachother. I will be in the chat as well so that I can direct you. I will tell you when to begin each round, what your MNS is for each round and who should make the initial request. You will negotiate in English.

Figure 1: Instruction sheet page 1



In order to reach a deal, you will need to take turns making requests. One player will make a request, and the other will say whether they agree to the request or make a "counter-request". If you request 6, for example, then the other player will receive 3. You may request as many or as few points as you wish to have. You can adjust your previous request to meet the other player's needs, or you can make the same request again if you think it is necessary. When a player agrees to a request, the round is immediately over and both players receive the agreed upon points. Each of you will take turns making the first request. To agree to an opponent's request, simply type "Deal." Then, I will announce to each player how many points he receives from the round, and how many points he has collected in total. This information will not be given to the other player.

You may also quit the round by typing "I quit." Once again, neither player will receive or lose any points if one player quits. Your MNS will not be subtracted from these zero points, so it's better to quit than to agree to a share under your MNS.

You will have 3 minutes to complete every round. I will give you a warning when 30 seconds are remaining and when 10 seconds are remaining. If you both cannot reach an agreement within 3 minutes, then neither of you receives or loses any points.

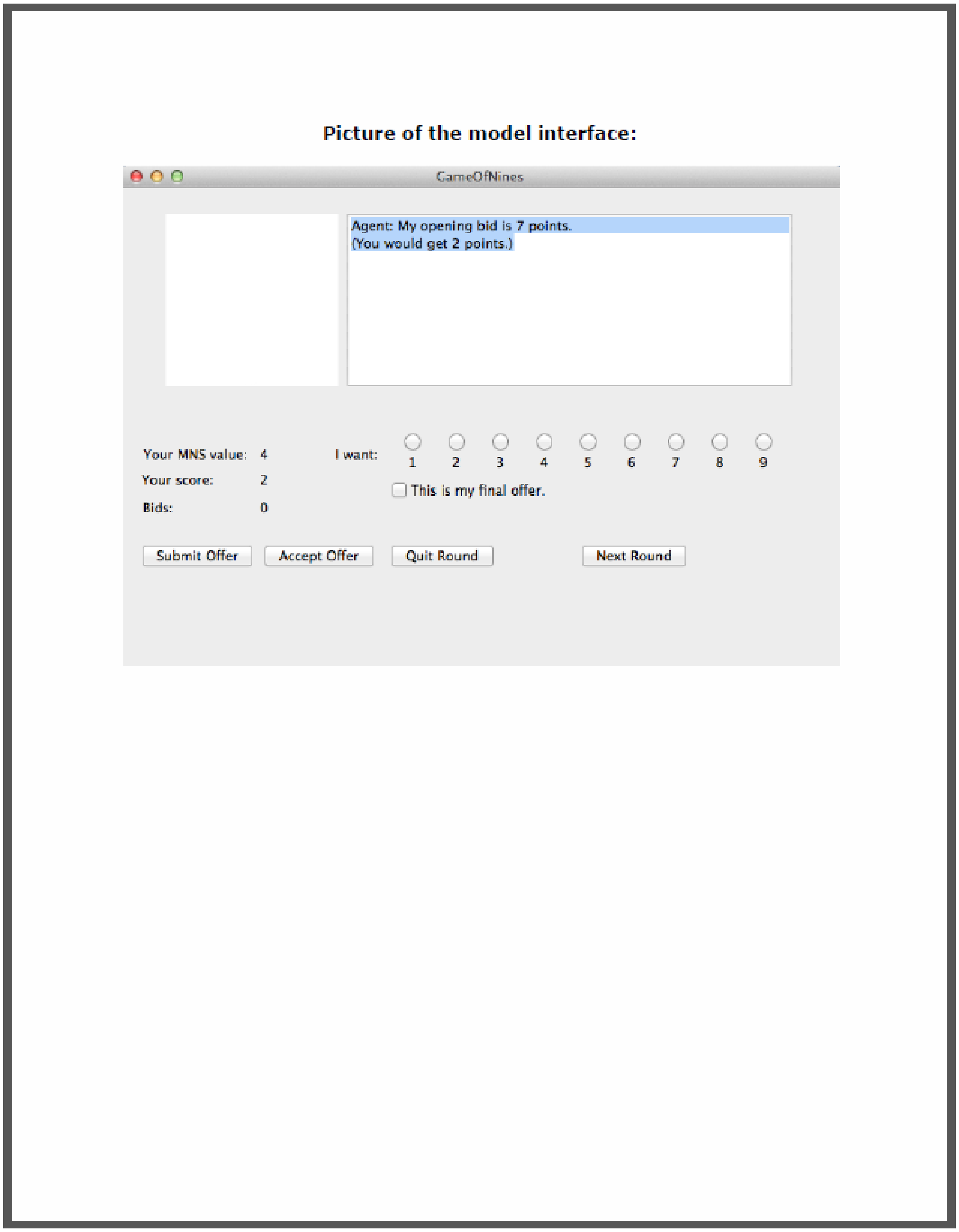
In this experiment, you will perform what is called a "Turing test". One player will play himself whereas the other player will either play himself, or he will use the actions of a computer model. The goal of the game is not to try to figure out if the other player is a human or a computer – trying to gain points is more important. You shouldn't use moves intended to figure out what the other player is. I will tell him what his role is once you are separated.

If you operate the model, you must always use the model's actions, both in the introductory and actual rounds. If you do not operate the model, you will be given a reaction time. You have to wait at least this time between receiving a message and sending a counter-message to emulate the model's slower response time.

Since the computer model cannot produce realistic messages, you'll use a set of predefined messages, which can be found in the message sheet. You may not use any other messages while negotiating.

At the end of the game, I will ask you to fill out a brief questionnaire. Then, I will announce the scores and the experiment will be over. If you have any questions, you may ask them before the experiment begins.

Figure 2: Instruction sheet page 2



Picture of the model interface:

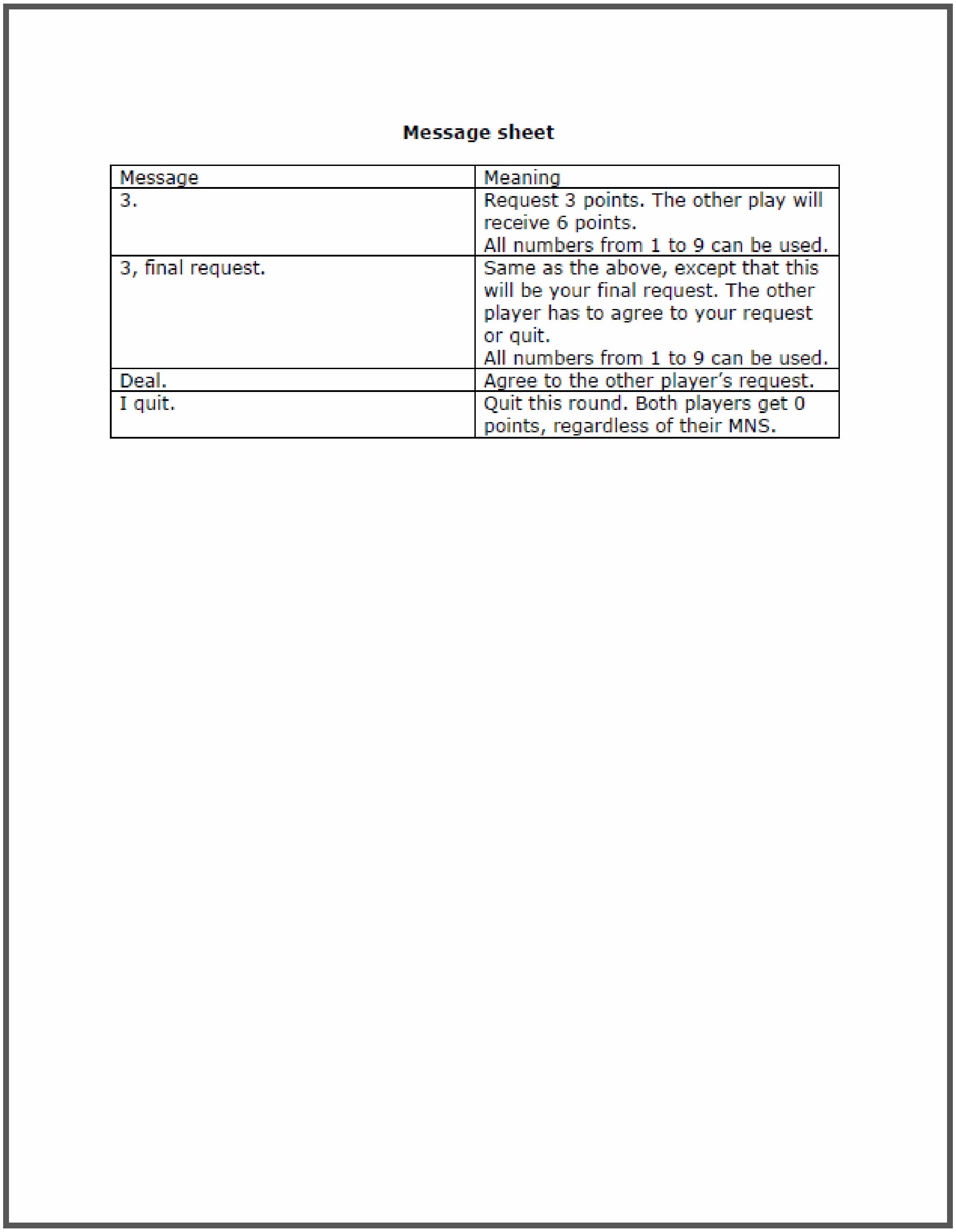
Agent: My opening bid is 7 points.
(You would get 2 points.)

Your MNS value: 4

Your score: 2

Bids: 0

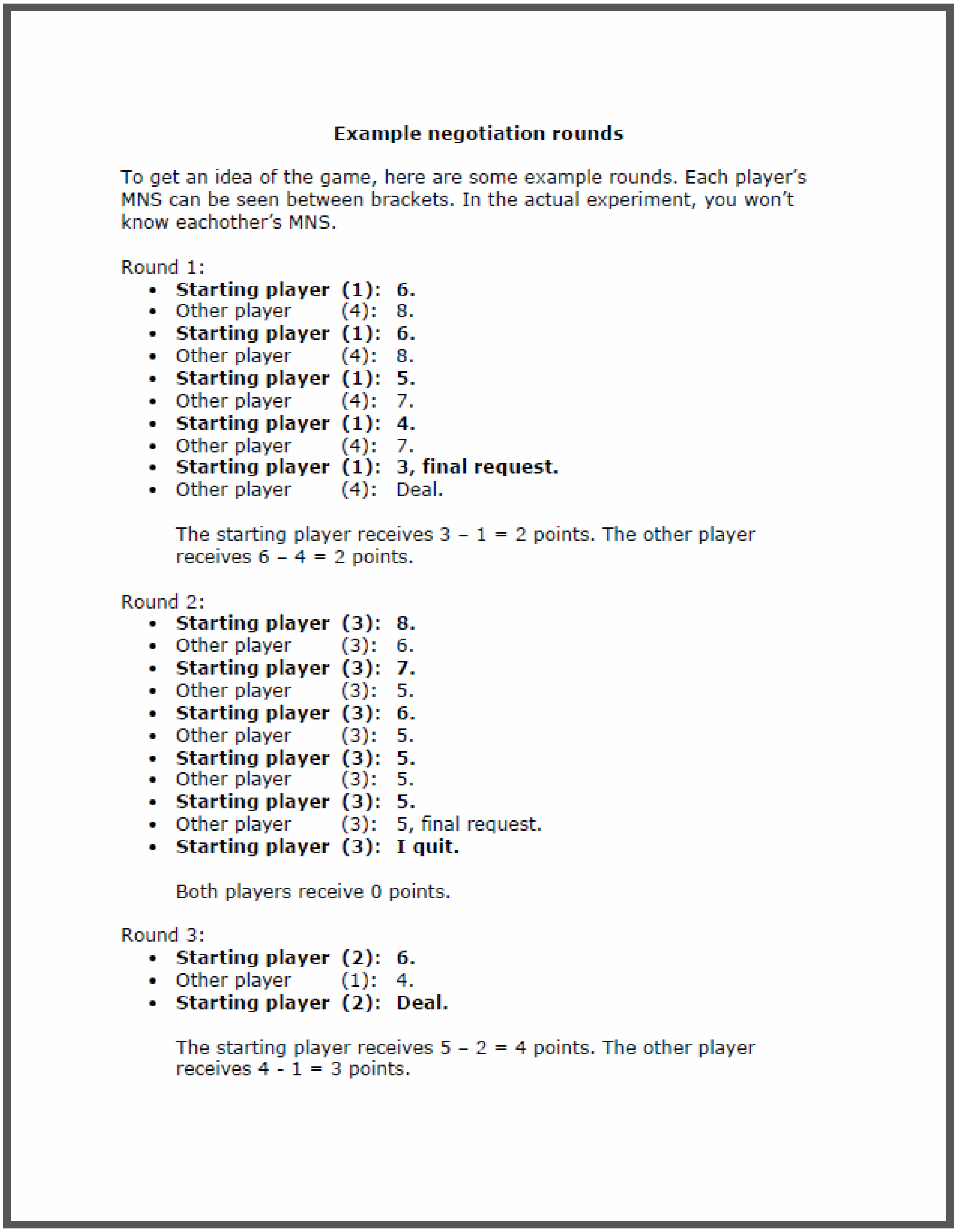
Figure 3: Instruction sheet page 3



Message sheet

Message	Meaning
3.	Request 3 points. The other play will receive 6 points. All numbers from 1 to 9 can be used.
3, final request.	Same as the above, except that this will be your final request. The other player has to agree to your request or quit. All numbers from 1 to 9 can be used.
Deal.	Agree to the other player's request.
I quit.	Quit this round. Both players get 0 points, regardless of their MNS.

Figure 4: Instruction sheet page 4



Example negotiation rounds

To get an idea of the game, here are some example rounds. Each player's MNS can be seen between brackets. In the actual experiment, you won't know eachother's MNS.

Round 1:

- **Starting player (1): 6.**
- Other player (4): 8.
- **Starting player (1): 6.**
- Other player (4): 8.
- **Starting player (1): 5.**
- Other player (4): 7.
- **Starting player (1): 4.**
- Other player (4): 7.
- **Starting player (1): 3, final request.**
- Other player (4): Deal.

The starting player receives 3 – 1 = 2 points. The other player receives 6 – 4 = 2 points.

Round 2:

- **Starting player (3): 8.**
- Other player (3): 6.
- **Starting player (3): 7.**
- Other player (3): 5.
- **Starting player (3): 6.**
- Other player (3): 5.
- **Starting player (3): 5.**
- Other player (3): 5.
- **Starting player (3): 5.**
- Other player (3): 5, final request.
- **Starting player (3): I quit.**

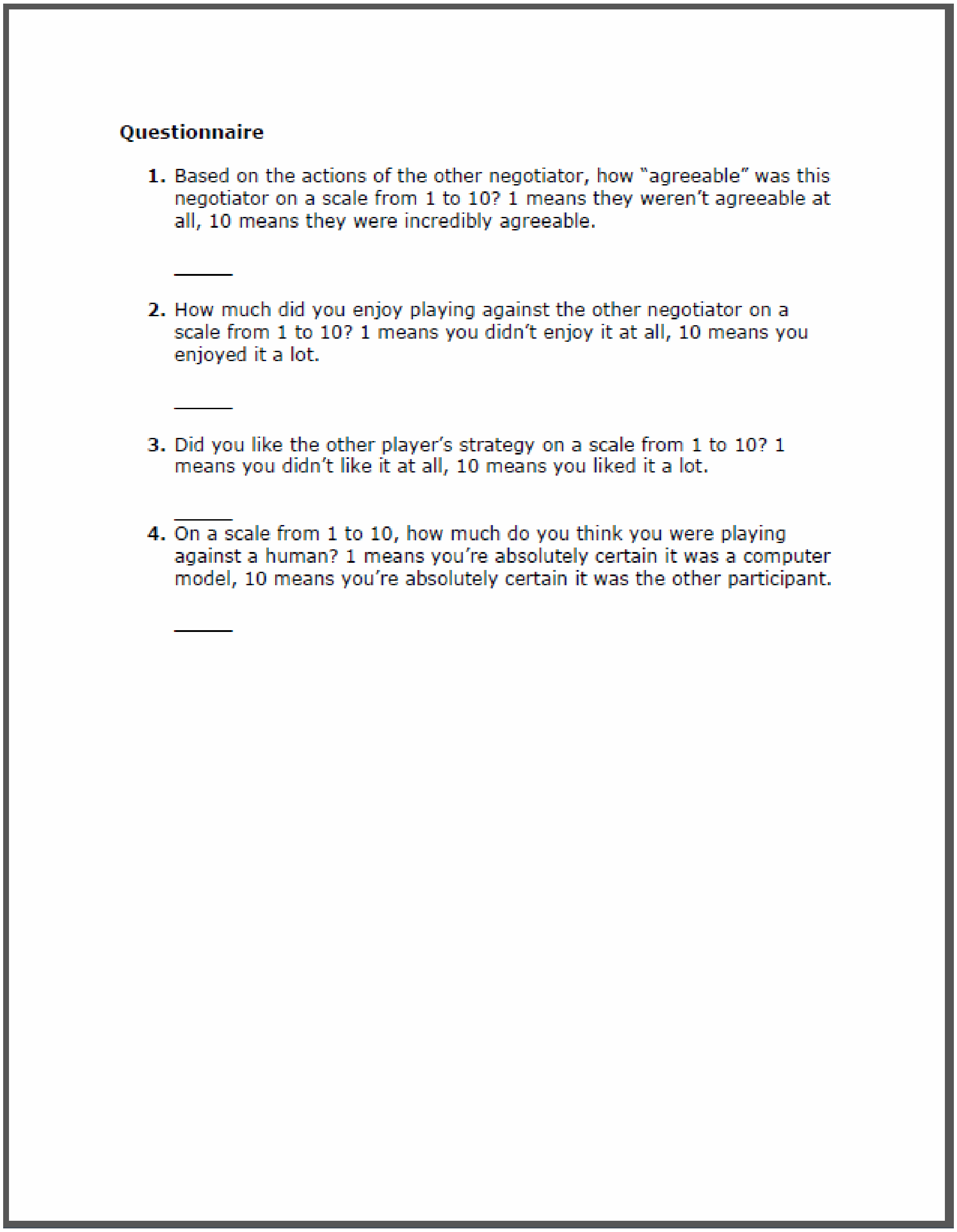
Both players receive 0 points.

Round 3:

- **Starting player (2): 6.**
- Other player (1): 4.
- **Starting player (2): Deal.**

The starting player receives 5 – 2 = 4 points. The other player receives 4 - 1 = 3 points.

Figure 5: Instruction sheet page 5



Questionnaire

1. Based on the actions of the other negotiator, how "agreeable" was this negotiator on a scale from 1 to 10? 1 means they weren't agreeable at all, 10 means they were incredibly agreeable.

   ______

2. How much did you enjoy playing against the other negotiator on a scale from 1 to 10? 1 means you didn't enjoy it at all, 10 means you enjoyed it a lot.

   ______

3. Did you like the other player's strategy on a scale from 1 to 10? 1 means you didn't like it at all, 10 means you liked it a lot.

   ______

4. On a scale from 1 to 10, how much do you think you were playing against a human? 1 means you're absolutely certain it was a computer model, 10 means you're absolutely certain it was the other participant.

   ______

Figure 6: Instruction sheet page 6