G-COE GLOPE II Working Paper Series

Strategic Sophistication Category: Response Time, Eye Movements and Stated Beliefs

Ai Takeuchi

Graduate School of Economics, Waseda University, JSPS Research Fellow (DC)

Yukihiko Funaki

Waseda University

Jana Vyrastekova

Radboud University, Department of Economics, Nijmegen,

Tilburg University, Department of Economics, Tilburg

Working Paper No. 11

Strategic Sophistication Category: Response Time, Eye Movements and Stated Beliefs

Ai Takeuchi[a]*, Yukihiko Funaki[b] and Jana Vyrastekova[c]

[a] Graduate School of Economics, Waseda University, JSPS Research Fellow(DC). 1-6-1, Nishi-Waseda, Shinjuku-ku, Tokyo 169-8050, Japan.

[b] Waseda University. 1-6-1, Nishi-Waseda, Shinjuku-ku, Tokyo 169-8050, Japan.

[c] Radboud University, Department of Economics, Nijmegen, The Netherlands. Tilburg University, Department of Economics, Tilburg, The Netherlands.

Abstract

This paper introduces a simple experimental tool, Strategic Sophistication Category (STS Category), designed to categorize subjects by whether they do or do not infer others' actions in strategic situations. We validate the design of STS Category by a novel mix of data on the subjects' response time, eye movements and stated beliefs in the STS Category. We found that subjects who have lower ability to infer others' action had shorter response time, looked less at the others' payoff matrix and had dispersed belief compared to those who were able to infer others' decisions.

Key Words: Strategic Inference, Matrix Game Experiment, Eye movements

JEL CODE: C72; C91

1 Introduction

In game theory, it is usually assumed that players have the ability to conduct infinite depth of strategic reasoning. That is, players are able to infer others' actions, infer how the others will infer his/her actions, infer how the others will infer how he/she infer others' actions and so on to infinity. However, it has been repeatedly observed in experiments that the number of steps of iterated dominance human

*Corresponding Author: Tel:+81 3 3442 8827, +81 3 3442 8827. E-mail Addresses: ai-tak@moegi.waseda.jp (A. Takeuchi), funaki@waseda.jp (Y. Funaki), J.Vyrastekova@fm.ru.nl (J. Vyrastekova)

subjects can perform is limited and varies across subjects. Individuals often conduct two to four steps of iterated dominance, with the median number of steps equal to two (see chapter 5 of Camerer, 2003). Moreover, a non-negligible proportion of subjects in the previous experiments on strategic sophistication,[1] is classified as non-strategic and not inferring others' action before making decisions.

This limited cognitive capacity and heterogeneity among human subjects needs to be accounted for when interpreting behavior observed in experimental studies and in the field. The theoretical benchmark for analyzing strategic interactions is usually based on the assumption of unlimited cognitive capacity in the sense of ability to infer actions of others. However, in many cases, the assumption of limited ability to infer actions of others on the side of the players might result in a different benchmark. In such situations, it becomes crucial to be able to distinguish the strategic sophistication of the subjects, and to control for this information in interpreting their behavior. To give an example, there is a discussion in the literature on the origins of sanctioning in the finitely repeated public good games where such sanctioning represents a secondary public good.[2] The discussion evolves around the possible strategic vs. nonstrategic reasons to supply sanctions. Besides using sophisticated experimental designs to address this question, one might also investigate the origins of sanctioning directly, by measuring the strategic sophistication of the sanctioning and nonsanctioning subjects.[3]

In this paper, we introduce a simple experimental tool designed to categorize experimental subjects into several categories by their ability to infer behavior of others. We refer to this tool as Strategic Sophistication Category (STS Category). Due to its no-feedback design, discussed below, STS Category is suitable as an auxiliary pre-experiment providing information on strategic sophistication of the participants, without spillovers for the main experiment performed by a researcher.[4]

STS Category consists of a series of two-person normal form games, and is based on the design by Costa-Gomes et al. (2001). The games used in STS Category are all strictly dominance solvable in pure strategies, thus always contains a unique Nash equilibrium. The classification of subjects based

[1]For example, there are many literatures which classify the subjects into different behavioral types in normal form games: Stahl and Wilson (1994, 1995), Costa-Gomes et al. (2001), Holt (1999), Rey-Biel (2008) and Costa-Gomes and Weizsäcker(2006) analyze the normal form games in matrix form; Nagel (1995), Ho et al. (1998) and Costa-Gomes and Crawford (2006) analyze the Beauty Contest Games.

[2]For example, see Fehr and Gachter (2000), Falk et al. (2005), Vyrastekova et al. (2008), Abbink et al. (2004) and Casari and Luini (2006).

[3]For an example of this approach, see Takeuchi (2007).

[4]Other examples of such auxiliary experiments that are used to supplement the researcher's main experiments are Holt and Laury (2002) instrument which allows an analysis between observed behavior and risk attitude (see Eckel and Wilson, 2004 for application) or Social Value Orientation which allows an analysis between the behavior and preference types (see for example, Offerman et al., 1996 for application).

on their choices in this sequence of dominance solvable games differs from the previous classifications mainly in the following two aspects. (1) First, we focus on classifying subjects by the difference in their ability to conduct first step of inference. In this way, we aim to capture strategic inference ability of an individual, which is less sensible to changes in the game structure than the categorization in the previous experiments.[5] Such stability of the experimental tool we introduce, even at the cost of specifying specific types in details, is to the benefit of the use of the tool across various experiments. (2) Second, since there is ample evidence which shows that subjects in experiments might not maximize their own monetary payoff, the design of STS Category separates subjects who seem to maximize their own monetary payoff apart from those who seem not, and only classify the former by their strategic sophistication[6]. This allows us to reduce the error in classification due to misinterpretation on behavior of the other regarding subjects. We discuss about the classification and the experiment design in detail in the following section.

The main contribution of this paper is the validation of the proposed STS Category by the use of additional sources of data, some of which has been previously found in the literature to correlate with the strategic sophistication of subjects. In particular, in our experiment on STS Category we measured, in addition to the behavior in the games, the subjects' response time, their eye movements, and we also included a treatment where subjects reported their beliefs on the actions chosen by other players in STS Category. We discuss the relationship of this data to strategic sophistication in Section 3.

We find, in accordance with expectations, that subjects who were classified as non-strategic had shorter response time, looked less at the others' payoff matrix and had more dispersed beliefs than subjects who were classified as strategic. These results suggest that STS Category is a valid tool for classification of the subjects by their strategic inference ability. Also, it indicates the possible effect of strategic sophistication on the behavior, since it affects the primitive cognitive data, such as response time and eye movements, as well as their beliefs of the others' actions.

The rest of the paper will proceed as follows. In the next section, we define STS Category, discuss its method of classification, and present the experiment design. In section 3, we review the literature

[5]If we compare the results of existing literatures, the percentage of each behavioral types in the population differs greatly between the experiments, even if they are using the same kind of behavioral types for the analysis. For example, even though the experiments in Stahl and Wilson (1994), Costa-Gomes et al. (2001) and Rey-Biel (2008) uses two person asymmetric normal form games, the results of these three papers differ greatly.

[6]Holt (1999) and Costa-Gomes et al. (2001) also had included types which are not maximizing monetary payoff in their analysis.

on measures related to the strategic sophistication of subjects, in particular the response times, eye movements and stated beliefs on behavior of others. Section 4 discusses the experiment conducted to validate STS Category, and Section 5 concludes.

2 Strategic Sophistication Category

This section explains the categorization of the subjects and the experiment design which classifies the subjects by their ability to infer others' choices. It allows us to check the existence of subjects who do not infer others' actions in making decisions. Also, it can serve as auxiliary experimental treatment that one can conduct along with their experiment to analyze the relationship between the strategic inference ability of an individual and the observed behavior.

The experiment is a series of two person normal form games in matrix form. All the games are dominance solvable in pure strategies and are solvable by either 1, 2 or 3 steps of iterative elimination of dominated actions[7] for both player role. Unlike the convention of the game theory, we count the number of steps of iterative elimination of dominated actions from the point of view of each player role, the row and the column. For example, if there is a dominant action for the row player but not the column, and if the column player have a dominant action in the reduced game after the elimination of the dominated action of the row player, than this game is solvable by 1 step of iterated dominance for the row player and 2 steps for the column player. The category uses the characteristics of the dominance solvable games and the results of the existing literatures on behavioral types. We have three major assumptions behind this classification.

First, as we had stated in the introduction, we assume that there are subjects who do and do not conduct first step of strategic inference.

Second, there are subjects who simply maximize their own monetary payoff (let us call them *MPM*, monetary profit maximizing, subjects) and those who also care about the payoff of the others (*Non-MPM* subjects). One of the most simplest example which show that some subjects in experiments are non-MPM are the observations in the dictator games. Even though the dictator is given a pie of a certain size that s/he can distribute as freely as s/he wishes, many give a positive portion of the pie to the receiver. Many other experimental evidences suggest that there are MPM as well as non-MPM

[7]Throughout this paper, we refrain from using the term "strategy" and instead use the term "actions."

subjects in experiments (see Camerer, 2003 for a survey.) At the same time, whether one is MPM or not tend to be condition dependent. For example, people are known to be sensible to intentions (Falk and Fischbacher, 2006, McCabe et al., 2003), social distance and entitlements (Hoffman et al., 1994, List and Cherry, 2008). Although we tried to design the treatment by raising the cognitive difficulty so it is more difficult for the individuals to become non-MPM, one cannot simply assume that all subjects are maximizing their own monetary payoff. This is one of the main reason why we define this category, instead of using models like the Cognitive Hierarchy Model (Camerer et al., 2004) where the assumption of homogeneity of preference is required.

Finally, we assume that when the subjects infer others' choices, they have a belief such that the others' preferences are like themselves. So, if a subject is MPM (Non-MPM) and can infer others' choices, than s/he believes that the others are also MPM (Non-MPM). One may think that if subjects are smart enough to infer others' choices, they should realize that there are some people with different preference than themselves. However, in the experiment by Costa-Gomes et al. (2001), no subjects were classified as being sophisticated enough to best respond to the distribution of the actual choices. Also, there is a tendency that when people put themselves into others' shoes they tend to expect that others are like themselves. In other words, the subjects' guess on the number of others who chose an action tends to be higher for those who chose that action themselves than for those who did not chose that action. This effect is called the *(False) Consensus Effect*, and is widely observed in experiments both in psychology (for example, Mullen et al., 1985, Ross and House, 1977, Dawes and Mulford, 1996) and in economics (for example, Offerman et al., 1996, Selten and Ockenfels, 1998, Engelmann and Strobel, 2000, 2004, Iriberri and Rey-Biel, 2008).

Based on these assumptions, we have created the experiment design and the payoff matrices so as to classify the MPM apart from Non-MPM and random individuals, and classify the MPM by whether they conduct the first step of inference.

2.1 Experiment Design

The experiment is based on the design by Costa-Gomes et al. (2001). In the experiment, subjects make choices in 14 two person asymmetric matrix games. All the games are dominance solvable in pure strategy, so the Nash equilibrium will be unique and will coincide with the equilibrium of iterated elimination of dominated actions.

The experiment proceeds as follows. First, instructions and understanding test (see Appendix B) is given to the subjects. At the start of the experiment, subjects are anonymously divided into the row or the column player, and will stay in this role during the whole experiment. To ascertain anonymity and to treat all the games as one shot, adjusted stranger matching protocol is used; at the beginning of each round, each subject is randomly matched with another subject in a different player role. So, in each round, a subject will be anonymously matched with a new subject in the other player role, and simultaneously makes a decision. Their choice and the choice of the matched individual will determine the payoff the subject will earn in that round. Their choice nor the matched players' choice effect the matching in the proceeding rounds.

During their decision making, s/he will not learn their payoff nor the other players' decision until the end of all 14 games. In other words, no information feedback is given to the subjects until the end of the treatment. This is to control for the information of the other players and to analyze the games as first responses of the subjects. It also allows for a design where the subjects can make their decisions at their own pace. This is helpful because, as we will show in the Section 4.2, there is a large variance in the response time of individuals and might induce more error if the rounds proceed only after everyone has made their decisions.[8]

The adjusted stranger matching protocol and the no information feedback design results in constant belief of a subject across all rounds. Engelmann and Strobel (2004) showed that people exhibits (false) consensus effect under no information feedback condition, but quickly adjust their beliefs to the information provided explicitly. Therefore, giving no feedback help support our third assumption explained previously that when people infer others' decisions, they believe that others have same preference as themselves.

Next, let us explain the display of the matrix game. Figure 1 is an example of the screen shown to the subjects. The payoffs of the deciding subject are on the left of the screen and the payoffs for the matched subject is on the right. Upon display, the payoff matrices of the column player are transposed so that all subjects make their choices as if they are row player. Subjects are informed in the instruction of this rotation of the display. Therefore, they do not know whether they are in the role of row or the

[8] Some subjects would have to wait for a long time until others have made their decisions, and this might make them tired of the experiment. More importantly, the waiting time might indicate some information about other subjects.

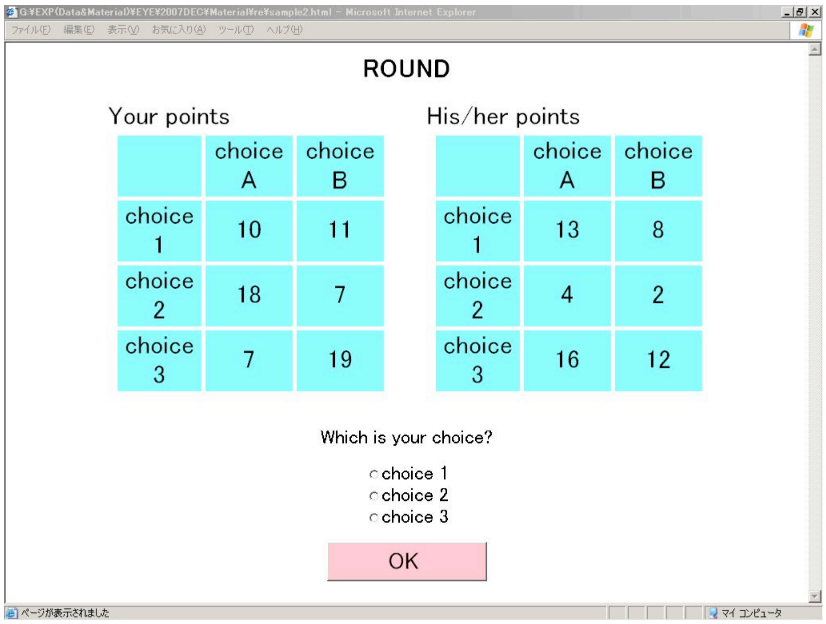


Figure 1: Example of a Screen Shown in the Experiment

column players. This allows us to control for the possible difference in decision making caused by being either a row or a column player.

Finally, at the end of the 14 games, subjects are given feedbacks on their choice and the payoff they had earned in all rounds. In the experiment, we referred to the payoff as points. Rather than making payment for the sum of all the points earned in each round, subjects are paid for one randomly chosen round to control for the wealth effect and to have to have the subjects make each decision seriously.

2.2 Categorization and Payoff in Games

This section explains the games used in the experiment along with the categorization of the subjects. These games are adjusted based on the games by Costa-Gomes et al. (2001) (see also Takeuchi, 2007), and are displayed in Table 1. As one can see, the number of actions for each player role is either 2 or 3 but there are no games where both players have 3 actions each. The games are asymmetric, and the numbers used for the payoffs ranged from 1 to 20.

In Table 1, just above each game, there are labels with two numbers in parenthesis. These numbers in parenthesis represent the number of steps of iterated dominance required to solve the game. All the games were dominance solvable and the combination of steps required to solve a game are either (1, 2), (2, 1), (3, 2) or (2, 3), so there are no games like prisoners' dilemma where both players have a

1A (1,2)

	Equilibrium	N, Mm, MM
Dominant	10 15	16 9
m-D,M-S	8 5	13 17

1B (2,1)

	Dominant	m-D, M-S
N, Mm, MM	10 17	18 14
Equilibrium	16 11	6 9

2A (1,2)

	Equilibrium		N, Mm, MM
Dominant	8 12	11 3	16 7
m-D, M-S	5 5	7 14	12 17

2B (2,1)

	m-D, M-S	Dominant
	15 8	4 12
Equilibrium	6 6	13 9
N, Mm, MM	18 13	8 17

3A (2,1)

	Dominant	m-D, M-S
Equilibrium	11 14	7 8
N, Mm, MM	8 11	18 9

3B (1,2)

	N, Mm, MM	Equilibrium
Dominant	10 7	13 10
m-D, M-S	8 17	7 6

4A (2,1)

	m-D	Dominant	M-S
N, Mm, MM	12 11	9 16	17 12
Equilibrium	7 5	14 10	8 4

4B (1,2)

	N, Mm, MM	Equilibrium
M-S	13 18	5 9
m-D	12 13	6 8
Dominant	17 8	11 15

5A (2,3)

	Dominated		
Equilibrium	4 5	10 14	15 8
N, Mm, MM	17 11	5 6	13 16

5B (3,2)

	N, Mm, MM	Equilibrium
Dominated	10 16	4 3
	12 15	7 14
	5 4	13 9

6A (2,3)

			Dominated
N, Mm, MM	11 18	10 6	19 13
Equilibrium	13 8	14 11	6 4

6B (3,2)

	N, Mm,MM	Equilibrium
	16 9	6 11
Dominated	11 17	2 4
	4 8	9 12

7A (1,2)

	N, Mm	Equilibrium	MM
m-D, M-S	4 18	8 3	16 19
Dominant	14 9	11 17	17 6

7B (2,1)

	m-D,M-S	Dominant
Equilibrium	4 9	18 12
MM	20 17	7 18
N, Mm	19 5	10 15

Dominant: Dominant Action
Equilibrium: Equilibrium of Iterative Elimination of Dominated Action
m-D: Action which minimizes the difference between the two players
M-S: Action which maximizes the payoff sum of the two players
N: Action corresponding to Naïve type (Best Responds to uniform distribution)
Mm: Maxi-min
MM: Maxi-max

Table 1: 14 Games To be Used in the Experiment

dominant action. For each player role, there are 5,7 and 2 games solvable by 1, 2 and 3 steps of iterated dominance, respectively.

Since the games are asymmetric, the row and the column player would be facing games with different payoffs. To be able to pool the data across the player role, half of the games are created by transposing the matrix, adjusting it by adding a small constant number (1 or 2) and by changing the ordering of the strategies. The labels on each games in Table 1 (such as 1A and 1B) indicate the games which are related in this way. Games labeled by the same number are created by this transformation. For example, in game 1A and 1B, the column player in 1B has the same payoff structure as row player in 1A. Therefore, the row and the column player in game 1A and 1B are playing games with same payoff structure just in a different order. This allows us to pool the data across the player role and analyze the row and the column players simultaneously.

The payoffs in the games are created to allow for the classification of (1) MPM subjects apart from Non-MPM or random subjects, and (2) MPM subjects who can and cannot infer others' decisions.

In order to classify the MPM subjects apart from Non-MPM or random subjects, we use the characteristic that, if the players are MPM, than the dominated actions are never best response given any belief of others' choices. Thus, given a game with dominant action, the MPM subjects will chose it, no matter what kind of belief they have about the choices of the other players. In games with dominant action, we created dominated action to correspond to the action that maximize the payoff sum of the two players and the action that minimize the payoff difference of the two players. So the action a MPM subject would chose (i.e. the dominant action) would not be chosen based on these two criterions (Max-Sum, Min-difference) which might reveal for the other regarding subjects' choice. Actions corresponding to these two criterions are indicated in Table 1 as M-S (Max-Sum) and m-D (Min-difference). We expect that Non-MPM subjects are likely to choose the dominated strategies while the MPM subjects will choose the dominant action. Therefore, we classify the subjects who do chose the dominant action for at least 80 % of the games with dominant action as MPM, and classify all the others as **U** (unclassifiable). We call this category unclassifiable because we cannot classify the strategic and non-strategic non-MPM subjects and also because we cannot classify the random player from non-MPM subjects from the choice data, and both will tend to be classified into this category.

To classify the MPM subjects by whether or not they infer others' actions, the characteristics of games solvable by 2 steps of iterated dominance can be used. If the subjects are MPM and can infer others' actions, then they will believe that the others will not choose the dominated actions because, from our third assumption, they believe that the others are also MPM. After the elimination of the others' dominated action, equilibrium action will be dominant in the reduced game. Therefore, MPM subjects who can conduct first step of strategic inference should chose the equilibrium action. As for the choice of the MPM subjects who do not infer the others' choice, we assumed the following three behavioral types based on the results of the existing literatures: Naïve, Maxi-max and Maxi-min. Naïve types best respond to the uniform distribution of the other player, Maxi-max chose the action which gives them the largest payoff and Maxi-min chose the action which maximizes the minimum payoff. These three behavioral types do not require strategic inference of others' choices; they can make their choice without looking at the others' payoff matrix. We have created the games solvable by 2 steps of iterated dominance so that the equilibrium action are separated from the actions these three behavioral types will chose (They are indicated in Table 1 by N (for Naïve) MM (for Maxi-max) and Mm (for Maxi-min)). Thus, we classify those subjects who do take the equilibrium action for more than 80% of the games solvable by 2 steps of iterated dominance as **H** (i.e., subjects with higher probability to infer the others choice) those that do take it for less than 20% are classified as **L**, and those in between are classified as **M**.

To summarize, let us define STS Category. For simplicity, let us call the ratio a subject took the dominant action in games solvable by 1 step of iterated dominance as D-Rate, and the ratio a subject took the equilibrium action in games solvable by 2 steps of iterated dominance as E2-Rate.

Definition 1. ***Strategic Sophistication Category (Takeuchi (2007))*** *A subject is in STS Category **U** (Unclassifiable) if the subject's D-Rate is less than 0.8. A subject is in STS Category **L** (Low) if the subject's D-Rate is greater than or equal to 0.8 and the E2-Rate is less than 0.2. A subject is in STS Category **M** (Middle) if the subject's D-Rate is greater than or equal to 0.8 and the E2-Rate is between 0.2 and 0.8. A subject is in STS Category **H** (High) if the subject's D-Rate is greater than or equal to 0.8 and the E2-Rate is greater than 0.8.*[9]

[9] One may wonder why games solvable by 3 steps of iterated dominance are included when only choices in games solvable by 1 and 2 steps are used in the classification. Actually, the subjects' choices in (2, 3) and (3, 2) games allow us to see whether the subjects are "best responding to a dominant action of the opponent" or "can also eliminate the dominated action of the opponent." There seems to be a difference between the two, latter being more difficult. Whether one can conduct the latter seems to be one of the main difference between the subjects in **H** and **M**, and adding games solvable by 3 steps of iterated dominance allows us to identify the subjects with higher strategic sophistication.

2.3 Discussion on the Thresholds of STS Category

Using the results of this experiment, one can classify the subjects into STS Category as follows. First, for each subject, count the number of times they took the dominant action in games 1A, 2A, 7A, 3B and 4B for row players and 1B, 2B, 7B, 3A and 4A for column players and divide it by 5, which is the number of games with dominant action. This is the D-Rate for each subject, and those with D-Rate less than 0.8 is classified as **U**. Similarly, for each subject, count the number of times they chose the unique Equilibrium action in games 3A, 4A, 5A, 6A, 1B, 2B, and 7B for row players and 3B, 4B, 5B, 6B, 1A, 2A and 7A for column players and divide it by 7, the number of games solvable by 2 steps of iterated dominance. This is the E2-Rate, and this value is used to classify those subjects who were not classified as **U**. If this value is less than 0.2, we classify them as **L**, if this value was greater than or equal to 0.8, than we classify them as **H**, and those that have values in between 0.2 and 0.8 are classified as **M**.

The threshold of 0.8 or 0.2 in the category corresponds to allowing for 1 error from each types' action, and is reasonable because it balances the two possible classification errors we like to consider. Let us denote the number of mistakes allowed in games with dominant action as k $(k = 0, ..., 5)$ and j $(j = 0, ..., 7)$ for the games solvable by 2 steps of iterated dominance. With these pairs of thresholds on the number of mistakes allowed in the classification, we can calculate the two possible classification errors. The first is the probability in which a type **H** player will not be classified as **H** due to small decision error. (The same probabilities can be calculated for type **L**. They will coincide with the probabilities calculated for type **H** due to symmetry in the definition.) We consider the case where, when a player is a certain type, s/he will take the action of the corresponding type, but will make a mistake with probability of ε and chose each action randomly. So in games with t actions ($t = 2, 3$), players take the action of their true type with probability of $1 - (t-1)\varepsilon/t$. Let us denote the probability of such error as $\bar{P}_{k,j}$. The second error we like to consider is the probability in which a random player who chose each action with a same probability will be classified as **H**, and let us denote it as $P_{k,j}$.

Table 2 shows these two possible errors in the classification for the pair of thresholds, where $\bar{P}_{k,j}$ is calculated with $\varepsilon = 0.1$. The calculation for this table is explained in more detail in Appendix C. As is clear from the definition, $\bar{P}_{k,j}$ decreases and $P_{k,j}$ increases as k or j increase. So there are no threshold that can minimize both probabilities and we need a threshold which balances the two possible classification errors. By comparing the two probabilities in Table 2, one can see that when 1 error is

Pair of Number of Errors Allowed in games solvable by 1 Step and 2 Steps (k and j)	$\bar{P}_{k,j}$	$P_{k,j}$
0 and 0	0.281	<0.001
1 and 1	0.021	0.005
2 and 2	0.001	0.067

$\bar{P}_{k,j}$: Pr(i is not classified as **H** $|i$ is **H** and $\varepsilon = 0.1$)
$P_{k,j}$: Pr(i is classified as **H** $|i$ is random)
Same probability can be calculated for subjects in **L** by replacing **H** by **L**.

Table 2: Thresholds and Probabilities of Classifying into Wrong Category

allowed for both games solvable by 1 step and 2 steps, these two probabilities are less than 5% and it balances the two probabilities.[10] Therefore, the threshold of 80 % (or 20 %) which corresponds to allowing for 1 error is a reasonable threshold to use for the classification of the subjects.

3 Validity Testing

This section tests whether STS Category is a plausible classification for those with and without the ability to infer the others choice. In other words, can we measure the strategic sophistication of MPM subjects by this simple experiment and check for the existence of subjects who do not infer others' actions? This section reviews the literature on response time, eye movement and stated beliefs to build hypothesis to validate the categorization.

3.1 Strategic Sophistication and Response Time

In Rubinstein (2007) it has been observed, using data sets with thousands of subjects, that those subjects who chose the action which seem to require more strategic reasoning, took longer time in making their decision.[11] So, if we measure the response time separately for each game, we can expect the response time of those subjects who were inferring the others' decisions to be longer than those subjects who do not infer the others' decisions. If STS Category successfully classify the subjects by their strategic

[10]This would not be the case if $\varepsilon \neq 0.1$. For each pair of thresholds, $\bar{P}_{k,j}$ would increase as ε increases.

[11]In Chong et al. (2005), they analyzed the relationship between the steps of thinking and response time. The steps of thinking were estimated using the Cognitive Hierarchy Model (Camerer et al., 2004). In their experiment, they did not find a clear relationship between the steps of thinking and the response time. In this paper, we will follow the result of Rubinstein (2007) in making our assumption for 2 reasons. First, if the subjects are attending to a more cognitive difficult task, it is likely that it is likely to take more time for them in making decision. Second, the sample size of the experiment by Chong et al. (2005) was 48, which is not as large as the sample analyzed in Rubinstein (2007). Large sample sizes are important because the response time data are very noisy, especially in problems like solving the dominance solvable games, where the subjects' talents may play a large role.

sophistication, the response time of subjects in category **L** should be shorter than that of subjects in **H**. Also, the response time of subjects in **H** can expected to differ among games with and without dominant action whereas it should not differ significantly for those subjects in **L**. This is because if the subjects infer others' decisions and are MPM, then the response time in games with dominant action should be shorter than those without, because there is no need to infer others' choices in games with dominant action. Whereas for those individuals who do not infer others' actions, the number of steps of iterated dominance required to solve a game should not effect their behavior.

Hypothesis 1. Subjects in **H** will have longer response time than subjects in **L**. Also, the response time of subjects in **H** will differ among games with and without dominant action, whereas not differ among subjects in **L**.

3.2 Strategic Sophistication and Eye Movements

In order for a subject to infer others' actions, it is necessary for them to look at the others' payoff matrix. On the other hand, for those subjects who do not infer others' actions, the information on their own matrix is enough to make a decision. Therefore, by measuring the eye movements of the subjects and analyzing the relative number of fixations on one's own matrix, we can analyze the validity of STS Category.

Eye movement analysis is a nice tool to analyze the higher cognitive process because it can detect and break down the eye movements of a person to less than 100ms for analysis, and are less likely to interfere with the subjects' higher cognitive process (Ohno, 2002). By analyzing the eye movements of the subjects, we should be able to have an idea of their thinking processes. In economics, the use of information search tracking and eye tracking has been increasing. For example, in relation to the strategic reasoning, Johnson et al. (2002), Costa-Gomes et al. (2001) and Costa-Gomes and Crawford (2006) tracked subjects' information search (using a software called Mouse Lab) and analyzed its relation with the decision making.[12] If we can analyze the relative information search on the two matrices, we can make following hypothesis. First, information on the others' payoff is necessary in

[12]Gabaix et al. (2006) also uses mouse lab to study choices among many goods under time pressure. Also, rather than the information search by the mouse, experiments which uses an eye tracking machine to analyze the relationship between the eye movement and behavior is increasing. Armel et al. (2008) studies choice behaviors, Innocenti et al. (2008) studies relationship between eye movement and information cascades, and Wang et al. (2008) analyze the eye movements in sender-receiver games.

order for an individual to infer others' actions, but if an individual do not infer others' actions, the information on their own payoff matrix is enough to make their decisions. Therefore, if STS Category classifies the subjects by their ability to infer the others' decisions, then the subjects in **L** should be looking relatively less at the others' payoff matrix than those subjects in **H**. Second, like with response time, **H** subjects' eye movements should be different in games with and without the dominant action, whereas it should be similar across these game categories for subjects in **L**. Even for those who do infer others' choices, it is not necessary for them to infer the others' decisions in games with dominant action, so the information on the others' payoff are not necessary, but for those who do not infer others' choices, their information search pattern should not differ among games.

Hypothesis 2. Subjects in **H** will look more at the others' payoffs than subjects in **L**. Also, the look up patterns of subjects in **H** will differ among games with and without dominant action, whereas not differ among subjects in **L**.

3.3 Strategic Sophistication and Stated Beliefs

By analyzing the subjects' stated beliefs, we can analyze the precision of the stated beliefs of subjects by looking at the variance of stated beliefs. If the subjects in category **L** do not infer others' choices, when we ask for the subjects to state their beliefs of the others' behavior, they are, in a sense, being forced to infer and make predictions of the other's behavior. Such tasks must be unfamiliar for subjects in **L** whereas more familiar for those subjects in category **H** who regularly infers other's choices. Therefore, the subjects in category **H** can expect to be able to state sharper beliefs of other's choices compared to subjects in category **L**.

Hypothesis 3. The stated beliefs of subjects in **L** will be more dispersed than that of subjects in **H**.

In order to test these hypothesis and to test for the validity of the category, we've conducted the following experiment.

4 Validity Testing Experiment

4.1 Experiment Design

The experiment was held on year 2006 and 2007 at Tilburg University's experimental laboratory CentER lab. The subjects were 114 students of Tilburg University whose majors were Economics, Law or Psychology. The language used in the experiment was English. The experiment was conducted using computers in isolated cubicles. Some cubicles were also equipped with an eye-tracking machine, which enabled us to measure the eye movements of some subjects.[13] Out of the 114 subjects, 53 subjects' eye movements were tracked. The technical details of the eye tracking can be found in the Appendix A.

The experiment had two treatments, the *STS treatment* and the *Guess treatment*. Each subjects participated first in STS treatment then in the Guess treatment.[14] The STS treatment is explained in the Section 2. The order of the games (see Table 1) in our experiment was as follows : 7B, 3A, 4A, 5B, 1A, 6A, 2A, 3B, 5A, 4B, 2B, 6B, 1B and 7A.[15] The subjects were paid for the payoff they earned in one of the 14 games, chosen randomly. The exchange rate was 0.25 Euros. Additionally, subjects were paid for their earnings from the Guess treatment (see below) plus the 5 Euro participation fee. The average earning for the STS treatment was 2.83 Euros.

After subjects participated in the STS treatment, the instructions for the Guess treatment were handed out, which they read on their own.[16] This was due to the location of the computers in cubicles and that we allowed individuals in the STS treatment to make their choices at their own pace. Since every subject ended at different timing, public reading was not plausible. We conducted an understanding test at the end of the instructions, and the subjects had to pass the understanding test in order to

[13] Due to the restriction of the number of eye tracking machines, after the normal sessions with 10 or more subjects each, additional sessions were held with just 4 subjects per session. The instructions used in the additional sessions had minor differences from the original sessions due to the limitation by the restricted number of subjects. First, to keep the decision situation anonymous, instead of being matched with the person in the same session as themselves, subjects were told that they would be matched with someone in the other group that had participated in the same experiment, not just from the subjects in the same sessions as themselves. Also, for the sessions with 10 or more subjects, the instruction for choice treatment were read out loud by the experimenter in a different room, but were read on their own for sessions with 4 or less subjects.

[14] The sessions with 10 or more subjects were conducted along with the treatment of VCM with Punishment stage, which we conducted to analyze the strategic and non-strategic motivations for sanctioning. The details of this treatment is available and are analyzed in Vyrastekova et al. (2008). We believe that this effect is small because it was done on a separate day and we stressed the point that two experiments were unrelated.

[15] In this experiment, the payoff matrices were slightly different from what we have in Table 1. In the original version, upon transforming the games (like from 1A to 1B), we multiplied the payoff by small amount before adding a constant, rather than just adding a constant. All the properties of the games are the same. The original games are available as the supplementary material on the website.

[16] Instructions used in the experiment is available as the supplementary material on the website.

proceed onto the Guess Treatment.

In the Guess treatment, the subjects were asked to state their beliefs on how many of the other subjects in their session chose each of the other's action available in a game. For belief eliciting, we used 6 games from the STS treatment (2A, 2B, 3A, 6A, 5B, 1A). Subjects were given a payoff matrix just like in the STS treatment, and instead of clicking on their choices, they were asked to state the number of person in the other group who chose A, B and C. The subjects' beliefs were given in whole numbers, and the numbers always had to add up to the number of subjects in the other player role. Therefore, depending on the number of subjects participating in the session, the numbers of subjects in the other player role changed. For example, if there were 12 subjects in one session, then the numbers assigned to each action had to add up to 6.[17] Subjects were paid for the accuracy of their guesses and the scoring rule we used for the payment was as follows:[18]

$$W - \alpha\big\{(A-a)^2 + (B-b)^2 + (C-c)^2\big\}$$

where A, B and C were the beliefs stated by the subjects and a, b and c were the number of subjects in the other player role who chose each of the actions. C and c were set to 0 in games where the opponent only had two actions. W was set to 20, and α was adjusted so the maximum points a subject could earn would be 20 and the minimum would be about 0 (α was set to 0.4, 0.3, 0.2 and 0.1 when the subjects in the session were 10, 12, 14 and 20 respectively.) The subjects were paid for 0.25 Euro times the number of points they had earned in one of the 6 rounds.[19]

[17]In the additional sessions we only had 4 subjects at a time. So instead of asking for the number of subjects in the other group who were in the same session as them, we had them state their beliefs of the frequency of action chosen by the 10 subjects whom were recruited the same way but will come in the other sessions.

[18]This scoring rule is different from the usual quadratic scoring rule used in, for example, Costa-Gomes and Weizsäcker(2006), because instead of asking for the probability the other player (whom they were matched with in the treatment where they made their choices) chose each of the action, we asked for their beliefs on the frequency of others who choose each action. The method we used, which is also used in Rey-Biel (2008) and Iriberri and Rey-Biel (2008), is intuitive and easy for the subjects to understand, but is not incentive compatible, although subjects will do no better by not stating their true beliefs. For the detailed explanation of the scoring rules, see Iriberri and Rey-Biel (2008).

[19]In the additional sessions conducted in 2007, the payment rate was set to 0.15 Euros per point, in order to raise the relative importance of the STS treatment. This seems to have slightly changed the stated beliefs of the subjects in some categories. We analyze for the possible incentive effect in footnote 27 and 28, after we introduce the variables we used for the analysis. The overall average earning for the Guess treatment was 3.75 Euros.

4.2 Data Analysis

In this section, we check for the validity of STS Category by analyzing its relationship with the response time, eye movements and stated beliefs. We first analyze the difference between the subjects in **L** and **H** to test for the validity of classification. Then look at the data of subjects in **M** and **U** at the end of the section to see the tendency of subjects in these two categories. The distribution of subjects used in this analysis were as follows: out of 114 subjects, 27 (23.7%) were categorized as **U**, 12 (10.5%) were categorized as **L**, 34 (29.8%) were categorized as **M** and 41 (36.0%) were categorized as **H**. Since we did not track every subjects' eye movements, the data used in the eye movement analysis are based on 53 subjects where 16 (30.2%) were classified as **U**, 7 (13.2%) were classified as **L**, 14 (26.4%) were classified as **M** and 16 (30.2%) were classified as **H**. So, even in our sample with mostly economic students, there did exist subjects who do not infer others' actions in our sample.

4.2.1 Comparison of L and H

First, let us start the analysis by looking at the difference in the response time of subjects in **L** and **H**. According to Hypothesis 1, if STS Category is successful in classifying the subjects, the response time of subjects in **L** should be shorter than that of subjects in **H**. Also, response time in games with dominant action should be shorter than other games solvable by either 2 or 3 steps for subjects in **H**, but not so different for subjects in **L**. In order to check these hypotheses, we measured the response time of each game in seconds. To increase the accuracy of the measurement, we included a preparation screen in between each game where we told the subjects to rest if they preferred in the preparation screen, but to concentrate during the actual decision making screen. The results are summarized in Figure 2 and Table 3. Table 3 are the mean and standard deviation of the response time of subjects in each category. It also shows the mean and standard deviation separately for the number of steps of iterated dominance required to solve a game. Figure 2 compares the empirical cumulative distribution function of response time of subjects in **L** and **H** for all games and also separately for games solvable by 1, 2 and 3 steps of iterated dominance.

First, from the left most graph in Figure 2, it is clear that the response time of subjects in **L** is shorter than that of **H**. The average response time of subjects in **H** is 36.14 whereas 20.63 for subjects in **L**, and we can reject the null hypothesis that the distribution of response time of subjects in the two categories

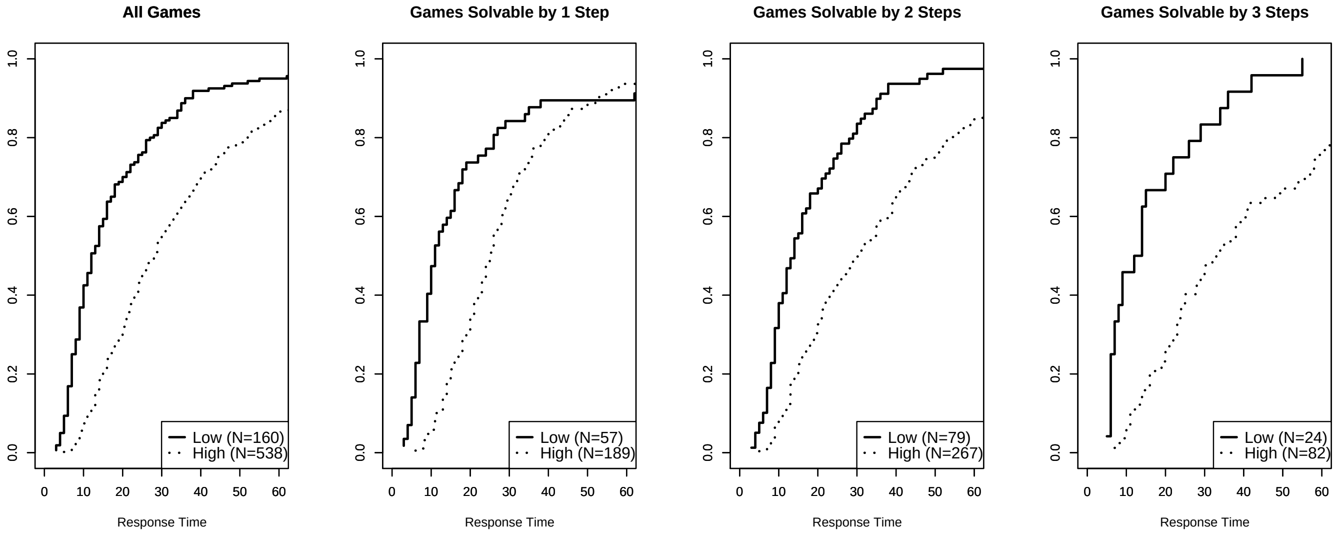


Figure 2: Comparison of the Empirical Distribution Function of Response Time of Subjects in **L** and **H**

	U	**L**	**M**	**H**
ALL Games	30.885	20.631	30.225	36.143
	(25.385)	(26.972)	(29.782)	(27.909)
Games Solvable by 1 step	31.156	23.649	25.205	30.360
	(27.544)	(36.861)	(20.047)	(21.525)
Games Solvable by 2 steps	30.115	19.557	33.607	37.536
	(23.434)	(21.021)	(35.902)	(27.914)
Games Solvable by 3 steps	32.870	17.000	30.603	44.939
	(26.962)	(13.500)	(24.513)	(36.973)
K. W. test (P-Value)^a	0.550	0.447	0.009	0.005

* Numbers in parenthesis are the standard deviation.
^{a} Results of Kruskal Wallis test with null hypothesis that the distribution of the response time is the same for all number of steps of iterated dominance required to solve a game.

Table 3: Comparison of Mean and Standard Deviation of Response Time for each Category

are the same at 1 % level (Wilcoxon rank sum test, W = 20831.5, p-value < 0.001). Next, by comparing the three graphs in Figure 2, one can see that the difference in the response time between subjects in **L** and **H** are smallest in games with dominant action, and is increasing with the steps of iterated dominance required to solve a game. Table 3 shows the p-values of the Kruskal Wallis test which tests, for each category, the null hypothesis that the distribution of response time is not different for the games with different number of steps of iterated dominance. From this we can see that the distribution of response time of subjects in **L** is consistent for all games, being unaffected by whether the game is solvable by 1, 2 or 3 steps of iterated dominance. On the other hand, for subjects in **H**, there was at least one distribution which was significantly different from other games solvable by different number of steps of iterated dominance. Results of the pair wise test showed that for the subjects in **H**, the difference in the distribution of response time in games solvable by 1 step and 2 steps, and the difference between 1 steps and 3 steps were both significant at 5 %, but were not significant between games solvable by 2 steps and 3 steps (pair wise Wilcoxon rank sum test using the p-adjustment method of Holms. All the pair wise test conducted in this paper uses the p-adjustment method of Holms.)[20] The response time of subjects in **H** is shorter in games with dominant action than other games which requires the elimination of the other's dominated action to solve a game. These results support the hypothesis, and STS Category is in line with the response time of the subjects.

Next, we analyze the relationship between STS Category and eye movements. The data used for the analysis is the number of times a subject fixates on the payoff matrices. The idea is that subjects with higher level of sophistication need to fixate over both their own and the others' matrices in order to go through the process of inferring other's action. In this analysis, we will use a measure called the ratio of number of fixations (henceforth, RNF).[21] RNF is the number of fixations a subject spent on his/her

[20] Since we analyze each game as an observation, not all of our observation is independent. Therefore, we calculated the mean of all games for each individual and treated this mean as an observation. We can still reject the null hypothesis that the distribution of average response time of subjects in **L** (N=12) and **H** (N=41) is the same at 1% level using Wilcoxon rank sum test. However, when we analyze the difference in the average response time for different number of steps of iterated dominance by calculating, for each individuals, the average response time separately for each number of steps, the results changed. For both subjects in **L** and **H**, we could not reject the null hypothesis that distribution of response time is not different for the games with different number of steps of iterated dominance using the Kruskal Wallis test. The p-values are 0.643 and 0.164 respectively. When we look closely at the data, there is difference in the distribution of response time for subjects in **L** and **H**. For subjects in **L**, the response time is about equal for all 3 kind of games, whereas, about 1/3 of the subjects in **H** decrease their response time slightly in the games solvable by 2 steps or 3 steps compared to games solvable by 1 step when the rest of the subjects increase their response time significantly. We speculate that these 1/3 of the subjects are the ones who start inferences from the other players' payoff matrix, but will need more samples with eye movements to analyze this speculation.

[21] The Fixation Filter was set with fixation radius of 30 pixel and minimum duration of 100ms. We've excluded observations where the total duration of the fixations on both matrices were equal to 0. These data is missing due to the positioning of

matrix divided by the number of fixations a subject spent on both matrices. So, going back to Figure 1, it is the number of fixation on the left matrix divided by the number of fixations on both matrices. Clearly, this will take a value between 0 and 1, 0 in cases where the subject never looks at his/her own payoff matrix, increase as the subject looks relatively more at his/her own payoff matrix, taking a value of 1 in cases where the subject never looks at the others' matrix. RNF is a good measure to see which matrix the subject was relatively focused on.[22] Using RNF, we can restate our two hypotheses in Hypothesis 2 as follows. Since subjects in **H** are supposed to be inferring the other's actions, they must fixate more on the other's payoff matrix (resulting in lower RNF) compared to subjects in **L**. Also, like in response time, the RNF in games with dominant action should be higher than other games solvable by either 2 or 3 steps of iterated dominance for subjects in **H**, but there should be no such difference for subjects in **L**. To test these hypothesis, we calculated the RNF of each subjects for each game and the data are summarized in Table 4 and Figure 3

Let us begin by checking the first hypothesis. From the comparison of empirical cumulative distribution function of RNF of subjects in **L** and **H** in all games (left most graph in Figure 3), one can see that the distribution of RNF is higher for subjects in **L** than **H**. The average RNF were 0.499 and 0.687 respectively, and we can reject the null hypothesis that the average RNF in all games are the same for the subjects in category **H** and **L** at 1 % level ($t = 8.271$, p-value < 0.001). Those subjects in category **L** looked relatively less at the others' payoff matrix compared to the subjects in **H**. Next, we compare within each category, the difference in RNF for the different number of steps of iterated dominance. From Table 4 and also from the graphs in Figure 3, one can see that RNF for subjects in **L** is consistent

subjects' head. Since the eye-tracking machine we used did not have any chin rest or head rest and allows for really natural data collection, sometimes, subjects move their head to the position where their eyes cannot be tracked by the machine with high precision. Since it is highly unlikely for a subject to make decision without fixating on one of the matrices, we excluded these observations. In the data used for this analysis, the minimum of the total duration in a game was 179ms.

[22]Using the eye movement data of the subjects, we can also analyze the number of switches from a matrix to matrix. The average (and the median in parenthesis) number of switches were 16.82 (12), 9.15 (5), 18.10 (12) and 13.08 (8) for subjects in **U**, **L**, **M**, and **H** respectively. There is significant difference in the distribution of number of switches between four categories, all the pair wise Wilcoxon test between each category were significant at 1% level, except for the difference between **U** and **M**. So, subjects in **U** and **M** had looked back and forth between the two matrices more than the subjects in **L** and **H**. Although this result is interesting, one must be careful in interpreting the data. Unlike the RNF, the number of switches does not control for the difference in the calibration between subjects. If there are two subjects with same eye movements but where one is better calibrated than the other, the number of switches for the better calibrated subject will be higher than the other. Also, for each game we can analyze subjects' first fixation. For each subject, we calculated the relative frequency of games they had first fixated on their own matrix. This value will be 1 for those who had always started looking from their own matrix whereas 0 for those who always start by inferring others choice. This value's average was 0.74(0.16), 0.80(0.13), 0.82(0.14) and 0.59(0.23) for subjects in **U**, **L**, **M**, and **H** respectively. There were 1 (out of 16) subject in **U** and 7 (out of 16) subject in **H** whose value was less than 0.5. So about half of the subjects in **H** seems to had higher tendency to start their decision making by inferring others' decisions rather than looking at their own payoff matrix.

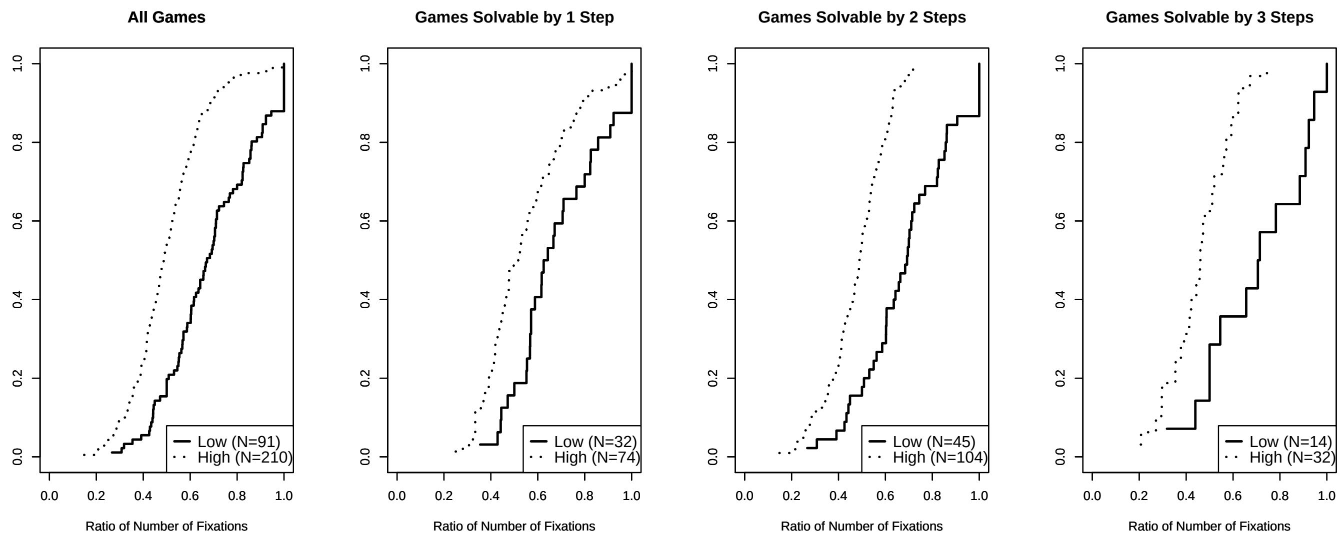


Figure 3: Comparison of Empirical Distribution Function of RNF of Subjects in **L** and **H**

	U	**L**	**M**	**H**
ALL Games	0.506	0.687	0.594	0.499
	(0.148)	(0.193)	(0.138)	(0.151)
Games Solvable by 1 step	0.513	0.680	0.599	0.545
	(0.124)	(0.186)	(0.137)	(0.179)
Games Solvable by 2 steps	0.502	0.687	0.600	0.479
	(0.159)	(0.194)	(0.141)	(0.126)
Games Solvable by 3 steps	0.506	0.702	0.566	0.457
	(0.161)	(0.216)	(0.131)	(0.132)
K. W. test (P-Value)[b]	0.907	0.906	0.729	0.067

[*] Numbers in Parenthesis are the standard deviation.
[b] Results of Kruskal Wallis test with null hypothesis that the distribution of the RNF is the same for all number of steps of iterated dominance required to solve a game.

Table 4: Comparison of Mean and Standard Deviation of RNF for Each Category

across games with different number of steps of iterated dominance, whereas decreasing for subjects in **H**. The results of Kruskal Wallis test shows that at least one distribution of RNF is different for some number of steps of iterated dominance for subjects in **H**, and the pair wise t-test shows that the difference in RNF between games solvable by 1 step and 2 steps and the difference between games solvable by 1 step and 3 steps are both significant at 5 % level (P-values were 0.010 and 0.011 respectively).[23] So the subjects in **H** are looking relatively more at the others' payoff matrix in games which requires elimination of the others' dominated action to solve the game, than in games with dominant action. These results suggest that STS Category is also in line with the eye movements of the subjects.

Finally, let us analyze the data from the Guess treatment and look at the relationship between STS Category and the stated beliefs of the individuals. In this section, we analyze the following two aspects of subjects' guesses; the precision of the prediction and the best response rate to their own stated beliefs. In both analysis, we will use the subjects' probability distribution implied by the stated number of other subjects in order to control for the size of the session. That is, if a subject's stated beliefs were N_A, N_B, and N_C in the raw data, the stated beliefs we will use for the analysis is P_A, P_B and P_C where $P_i = N_i/(N_A + N_B + N_C),\ \ i = A, B, C.$

First, before looking at the precision of the beliefs, we check whether the subjects believed that other subjects will choose the dominant action or the equilibrium of iterated domination. The average relative belief placed on the dominant action were 0.62, 0.65, 0.78 and 0.87 for subjects in category **U**, **L**, **M** and **H** respectively. So all subjects seems to have believed that relatively high portion of the others will choose the dominant action. (For the subjects in **H**, in 70% of the cases they stated that all subjects in other player role would choose the dominant action. This supports our third hypothesis in Section 2.2 that if the subjects will infer others' actions, they would expect that the others to have the same preference as themselves.) The relative belief on the equilibrium action decreases as the number of steps of iterated dominance increases. The mean relative belief on the equilibrium action in games solvable by 2 steps of iterated dominance were 0.160, 0.194, 0.275 and 0.585 and were 0.222, 0.083, 0.176 and 0.439 in games solvable by 3 steps for subjects in **U**, **L**, **M** and **H** respectively.

[23] To deal with the problem of independence, we calculated the average RNF for each individuals across games and compared the RNF of subjects in **L** (N=7) and **H** (N=16). The result was still significant at less than 1% level using the t-test. In RNF, the results also hold for the comparison of the RNF in games with different number of steps of iterated dominance within each category. The p-values of Kruskal Wallis test were 0.989 for **L** and 0.074 for **H**. The result of pair wise t-test for subjects in **H** showed that the difference between games with dominant action and games solvable by 3 steps of iterated dominance at 5 % level and the difference between games solvable by 2 steps of was almost significant at 10 % level with p-value of 0.112.

To measure how dispersed the subjects' stated beliefs are, we use the unbiased variance of the stated beliefs.[24] The variance of stated beliefs would be higher as the beliefs become more extreme, taking the maximum value 0.5 or 0.333 when the number of others' choices are 2 or 3 respectively, and would be lower as it become closer to uniform distribution, taking the minimum value 0 when it exactly matches it. As we had previously stated, inferring and stating beliefs of others' choices must be unfamiliar task for subjects in **L**. Therefore, we can expect their state beliefs to be more dispersed and closer to uniform distribution resulting in lower variance, compared to subjects in **H** (see Hypothesis 3).[25]

	U	**L**	**M**	**H**
N	27	12	34	41
Mean (Standard Deviation)	0.198 (0.192)	0.177 (0.184)	0.224 (0.188)	0.271 (0.194)
Median	0.124	0.092	0.16	0.222

Table 5: STS Category and Variance of Stated Beliefs

Table 5 lists the mean, median and standard deviation of the subjects' variance of stated beliefs. Both the mean and the median of variance of stated beliefs is lower for subjects in **L** than **H**, and we can reject the null hypothesis that the distribution of the variance of stated beliefs is the same at 1% level (Wilcoxon rank sum test, W=6267.5, P-value $<$ 0.001)[26]. We also calculated the variance of stated beliefs separately for the games where the other subject has 2 actions and games with 3 actions. In the games with 2 actions for the opponent, the mean (and the median in parenthesis) of the variance of beliefs was 0.210 (0.091) for subjects in **L** and 0.325 (0.5) for subjects in **H**. In the games where the other player has 3 actions, it was 0.110 (0.076) for subjects in **L** and 0.163 (0.146) for subjects in **H**. The difference in the mean and median is smaller for games with 3 actions, but we can still reject the null hypothesis that the distribution of the variance of stated beliefs is the same for the subjects in **L** and **H** at 5% level for both games with 3 actions and 2 actions (Wilcoxon rank sum test, P-value = 0.001

[24]So, if the subjects' stated belief for a game with 3 choices for the opponent were P_A, P_B and P_C, the variance of stated belief is $1/2 \times [(P_A - 1/3)^2 + (P_B - 1/3)^2 + (P_C - 1/3)^2]$, and $(P_A - 1/2)^2 + (P_B - 1/2)^2$ for games with 2 actions for the opponent.

[25]Notice that this definition of precision is different from accuracy. Being able to state precise belief is different from being able to state accurate beliefs, and in some situations depending on the distribution of the subjects, these two might even oppose each other. In our view, strategic inference ability of an individual is related to the precision in the reasons stated, but there is no clear relationship with accuracy. Even if individual infer others' choices, there is no guarantee that the result of his/her inference will always match the reality.

[26]However, this does not mean that subjects in **L** were choosing uniform distribution with a higher probability. We counted the number of times the subjects in **L** and **H** were stating beliefs which all actions were taken with probability between 0.3 and 0.35 or 0.45 and 0.55 in games with 2 or 3 actions respectively. Subjects in **L** stated such beliefs only 2 out of 72 possible times and subjects in **H** stated 13 out of 246 times.

and 0.020 respectively). Therefore, this result that the variance of stated belief is lower for subjects in **L** than **H** seem to be fairly robust.[27]

Next, let us analyze the frequency of actions being best responses to the same subjects' stated beliefs. In the two researches which are closely related to this, there was a large difference in the frequency of actions which were best responses to the stated belief. In the experiment of Costa-Gomes and Weizsäcker (2006) subjects best responded to their own stated belief in 51% of the games (7.08 games out of 14 games) on average, and in the experiment by Rey-Biel (2008), subjects' best responding rate were 69.25% in constant sum games and 66.13% in non constant sum games. However, in the two experiments, the most common behavioral types were different. In Costa-Gomes and Weizsäcker(2006), Naïve (L1) type was the most common behavioral type whereas in Rey-Biel (2008), Nash Equilibrium was the most common type, especially in the constant sum games. Since the former would be categorized as **L** and latter **H**, the frequency of best responses might be different between the subjects in **L** and **H**.

In order to check this, the number of games a subject was taking the best response to his/her own stated beliefs was calculated for each subject. Figure 4 show the proportion of subjects in category **L** and **H** for each number of best responded games. As we can see from this figure, subjects in category **H** tend to choose the best responce to their own beliefs more than those subjects in category **L**. The average number of games best responded to the stated beliefs were 4 for subjects in **L** and 4.902 for **H**, and we can reject the null hypothesis that the distribution of the number of best responded games are the same between **L** and **H** at 5% level (Wilcoxon rank sum test, W = 146, p-value = 0.026).[28]

4.2.2 Analysis of Category M and U

In this sub-section, we will look at the data of subjects classified as **M** and **U**. Here, we will use the response time and eye movement to make inferences about the relative frequency of the subjects who

[27] Since we changed the incentive of Guess treatment, we analyzed the difference in variance for the low (0.15 Euro per point) and high (0.25 Euro per point) incentive condition. The variances were lower in the low incentive condition (Wilcoxon Rank Sum test, $W = 56371$, p-value < 0.001). When we analyze the difference within each category, the difference ws significant at 1 % level for subjects in **M** and 5 % for **H**. Still, even in the low incentive condition, the variance was lower for subjects in **L** than **H** although this difference was slightly insignificant at 10 % (Wilcoxon Rank Sum test, $W = 643.5$, P-value = 0.118).

[28] Let us analyze the effect of low and high incentive in the number of games best responded. The number of games best responded were lower for low incentive condition (Wilcoxon Rank Sum Test, W =146, P-value = 0.022), but none of the difference within each category were significant at 5 % level. Still, the difference in the best responses of subjects in **L** and **H** disappeared when only using the low incentive condition data, probably because of the small sample size.

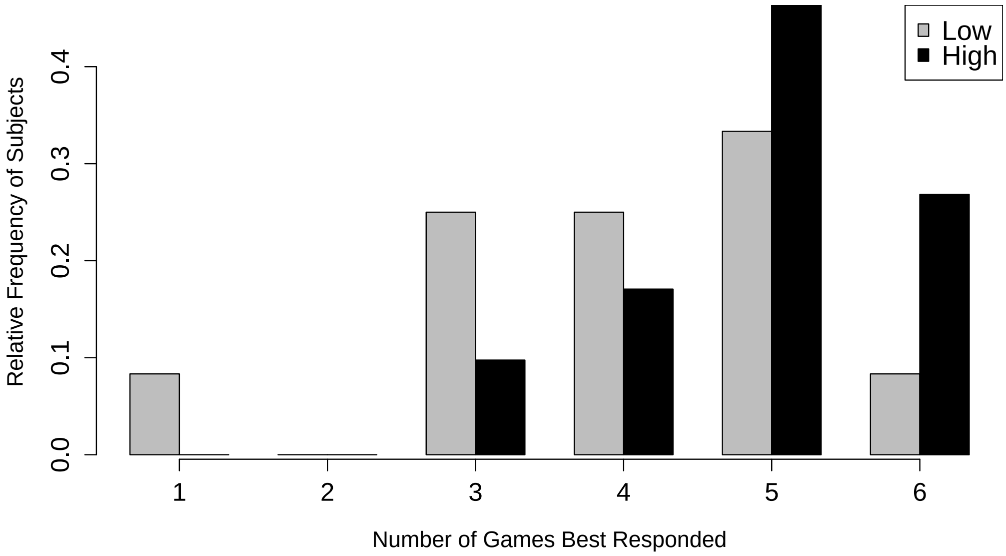


Figure 4: Comparison of Distribution of Number of Games Best Responded to the Stated Beliefs

chose randomly.

The analysis in this section is still tentative, since the response time and RNF could include a large error. Nevertheless, the results are interesting.

First, let us start with category **M**. From the definition of STS Category, the characteristic of **M** is not as clear as that of subjects in **L** or **H**. Intuitively, we can expect subjects in **M** to have characteristics in between subjects in **L** and **H**, but, if there are subjects who can find the dominant action but chose randomly whenever there is none, they are also likely to fall into this category. If there are many such subjects in category **M**, then their response time in games solvable by 1 step of iterated dominance should be longer than the games solvable by 2 or 3 steps, and the RNF should be lower. On the other hand, if those subjects in **M** have tendency in between those of **L** and **H**, then we can expect the data to lie in between that of **L** and **H**.

Going back to Table 3, we can see that the response time of subjects in **L** is lower than that of **M** which is lower than that of **H**. Pairwise Wilcoxon rank sum test was conducted, and all the pair wise comparison were significant at less than 1% level. Also, from Table 4 the mean RNF of the subjects in **M** was in between those of **L** and **H**. Pairwise Wilcoxon rank sum test was again conducted, and all the pair wise comparison were significant at less than 1% level. These results suggest that subjects who do take the dominant action but randomize when there is none, are not likely to exist, at least with a higher frequency, and the subjects in **M** are subjects with characteristics in between subjects in **L** and **H**.

Finally, let us look at the data for subjects in category **U**. Subjects in **U** can be a random player or a subject with a social preference. It is not possible to tell these two apart from the RNF or the stated beliefs. We can expect that those subjects with a social preference to look at both matrices for about the same time, and this tendency can be seen in the data set (i.e. we cannot reject the null hypothesis that the average of the RNF for the subject in **U** is equal to 0.5. The results of the t-test was $t = 0.587$ and p-value = 0.558). However, this same prediction can also be made for the players who chose randomly. Also, we cannot make any hypothesis on the variance of the stated beliefs for the subjects in **U**. However, the response time can be used to judge the type of subjects in category **U**. If the subjects in **U** are choosing randomly, then their response time should be shorter than the subjects in other categories. If the subjects in **U** are taking the other player's payoff into account and have some preference other than maximizing their own payoff, then their response time should vary with the number of the actions for her/himself and the other player, rather than with the steps of iterated dominance. If the subjects in **U** are a mixture of subjects who chose randomly and subjects with a social preference, then the standard deviation of the response time in **U** should be larger than the other types.

To compare the response time of subjects in **U**, look again at Table 3. The average response time of subjects in **U** is significantly larger than the response time of subjects in **L**, have no difference with **M**, and are significantly lower than that of **H**. The p-values of the of pair wise Wilcoxon rank sum test were $p < 0.001$, $p = 0.485$ and $p < 0.001$ respectively. This means that even though they are not inferring others' actions as much as **H**, they do not seem to choose randomly. Also, the steps of iterated doiminance had no effect on the response time of subjects in **U**, for there was no significant difference in the distribution of response time of each number of steps of iterated dominance (Kruskal-Wallis chi-squared = 1.197, p-value = 0.550). Although not significant, there was a positive relationship between the response time and the product of the number of actions of the two players. In the cases where one of the two had 3 actions and the number of cells equal 6, the mean response time was 32.2 whereas in cases where the number of cells equal 4, the mean response time was 27.843. Therefore, it is more likely that the subjects in Category **U** are not all random players, but include some players with a social preference. Now, if the subjects in **U** were a mixture, the standard deviation of the response time in **U** should be larger than other categories, since the response time of the random subject can expected to be short, whereas that of the subject with a social preference can expected to be long. However, the standard deviation of the response time in **U** is not so different from the other categories. Therefore, the data of the reponse time seems to suggest that subjects in category **U** are subjects with a social preference, and are not random players.

To summarize, the experiment data suggests that subjects in **U** are likely to be Non-MPM subjects and that subjects in **M** are MPM subjects who sometimes do and sometimes do not infer others' choices. However, this is just an result of one experiment and it is possible that we just had a lucky population with no or very few random players. If there is an individual who choose randomly, we could expect them to fall into one of the two categories with high probability.

5 Conclusion

We propose an experimental tool for categorization of subjects based on their ability to infer actions of others in strategic environments. Our measurement tool, referred to as strategic sophistication category (STS Category), allows to classify apart subjects who do and those do not infer others' actions. STS Category modifies the design by Costa-Gomes et al. (2001), simplifies the treatment of non-strategic behavioral types and extends it by accounting for the possibility that subjects hold other-regarding preferences. By filtering out such subjects, the error rate in classifying subjects as non-strategic because of their interpretation of the payoffs in the game is reduced.

In particular, STS Category allows to classify subjects into four categories, **U**, **L**, **M** and **H**, where **U** are subjects who either choose their actions at random or subjects with preferences other than maximizing only their own monetary payoff, **L** are subjects who do not infer others' actions, **H** are subjects who do infer others' actions and **M** are subjects in between **L** and **H**.

Our main result is the validation of STS Category by a novel mix of additional data, which has previously been linked in the literature to the strategic sophistication of subjects. This data consists of subjects' response time, eye movements during the decision-making, and stated beliefs about the behavior of others in the experiment.

In our experiments, we find about 10% of subjects to be classified as classified as **L** and 36% to be classified as **H** types, reflecting a heterogeneity of our subject pool in terms of ability to infer behavior of others.

More importantly, we find evidence validating STS Category. Subjects in category **L** had shorter response time, had looked relatively less at the others' payoff matrix and had more dispersed beliefs compared to subjects in **H**. Also, the response time and eye movements of subjects in **H** were sensitive to the necessity of inference; the response time was longer and they tend to look more at the others' payoff matrix in games solvable by two or three steps, compared to games with a dominant action where the inference was not necessary. Such tendency was not observed for subjects in **L**. Based on these observations, we conclude that STS Category is a valid tool that does classify the subjects by their strategic inference ability.

We argue for the use of STS Category in the role of an auxiliary experiment in studies where the level of ability to infer behavior of others might significantly affect the theoretical benchmark predictions of the researcher. We hope that by the use of STS Category, one will be able to interpret the behavior in strategic environments with a greater precision, and shed a new light on many aspects of strategic behavior.

References

Abbink, K., Sadrieh, A., Zamir, S., 2004. Fairness, public good, and emotional aspects of punishment behavior. Theory and Decision 57, 25–57.

Armel, K. C., Beaumel, A., Rangel, A., 2008. Biasing simple choices by manipulating relative visual attention. Judgment and Decision Making 3 (5), 396 – 403.

Camerer, C. F., 2003. Behavioral Game Theory: Experiments in Strategic Interaction. Princeton University Press.

Camerer, C. F., Ho, T.-H., Chong, J.-K., 2004. A cognitive hierarchy model of games. The Quarterly Journal of Economics 119, 861 – 898.

Casari, M., Luini, L., Apr. 2006. Peer punishment in teams: Emotional or strategic choice? Purdue University Economics Working Papers 1188, Purdue University, Department of Economics.

Chong, J.-K., Camerer, C. F., Ho, T.-H., 2005. Cognitive Hierarchy: A Limited Thinking Theory in Games. Springer, Ch. 9, pp. 203 – 228.

Costa-Gomes, M. A., Crawford, V. P., 2006. Cognition and behavior in two-person guessing games: An experimental study. The American Economic Review 96 (5), 1737 – 1768.

Costa-Gomes, M. A., Crawford, V. P., Broseta, B., 2001. Cognition and behavior in normal-form games: An experimental study. Econometrica 69, 1193–1235.

Costa-Gomes, M. A., Weizsäcker, G., 2006. Stated beliefs and play in normal form games, harvard University, mimeo.

Dawes, R. M., Mulford, M., 1996. The false consensus effect and overconfidence: Flaws in judgment or flaws in how we study judgment. Organizational Behavior and Human Decision Processes 65 (3), 201 – 211.

Eckel, C. C., Wilson, R. K., 2004. Is trust a risky decision? Journal of Economic Behavior and Organization 55, 447 – 465.

Engelmann, D., Strobel, M., 2000. The false consensus effect disappears if representative information and monetary incentives are given. Experimental Economics 3, 241–260.

Engelmann, D., Strobel, M., 2004. The false consensus effect: Deconstruction and reconstruction of an anomaly, cERGE-EI Working Paper Series.

Falk, A., Fehr, E., Fischbacher, U., 2005. Driving forces behind informal sanctions. Econometrica 73 (6), 2017–2030.

Falk, A., Fischbacher, U., 2006. A theory of reciprocity. Games and Economic Behavior 54, 293–315.

Fehr, E., Gachter, S., 2000. Cooperation and punishment in public goods experiments. American Economic Review 90 (4), 980–994.

Gabaix, X., Laibson, D., Moloche, G., Weinberg, S., 2006. Costly information acquisition: Experimental analysis of a boundedly rational model. The American Economic Review 96 (4), 1043 – 1068.

Ho, T.-H., Camerer, C. F., Weigelt, K., 1998. Iterated dominance and iterated best response in experimental 'p-beauty contests'. American Economic Review 88, 947–969.

Hoffman, E., McCabe, K., Shachat, K., Smith, V., November 1994. Preferences, property rights, and anonymity in bargaining games. Games and Economic Behavior 7 (3), 346 – 380.

Holt, C. A., Laury, S. K., 2002. Risk aversion an incentive effects. The American Economic Review 92 (5), 1644 – 1655.

Holt, D. J., 1999. An empirical model of strategic choice with an application to coordination games. Games and Economic Behavior 27(1), 86–105.

Innocenti, A., Rufa, A., Semmoloni, J., 2008. Cognitive biases and gaze direction: An experimental study, labsi Working Papers.

Iriberri, N., Rey-Biel, P., June 2008. Elicited beliefs and social information in modified dictator games: What do dictators believe other dictators do?, mimeo.

Johnson, E. J., Camerer, C. F., Sen, S., Rymon, T., 2002. Detecting failures of backward induction: Monitoring information search in sequential bargaining. Journal of Economic Theory 104, 16–47.

List, J. A., Cherry, T. L., 2008. Examining the role of fairness in high stakes allocation decisions. Journal of Economic Behavior and Organization 65, 1 – 8.

McCabe, K. A., Rigdon, M. L., Smith, V. L., 2003. Positive reciprocity and intentions in trust games. Journal of Economic Behavior and Organization 52, 267–275.

Mullen, Atkins B., Champion J. L., Edwards D. S., Hardy D. H., J. E. D. Story, M. Vanderklok, 1985. False consensus effect: A meta analysis of 115 hypothesis test. Journal of Experimental Social Psychology 21, 263 – 283.

Nagel, R., 1995. Unraveling in guessing games: An experimental study. American Economic Review 85, 1313–1326.

Offerman, T., Sonnemans, J., Schram, A., July 1996. Value orientations, expectations and voluntary contributions in public goods. The Economic Journal 106, 817 – 845.

Ohno, T., 2002. What can be withdraw from eye movement? - understanding higher cognitive processes from eye movement analysis. Cognitive Studies 9(4), 565 – 576, in Japanese.

Rey-Biel, P., 2008. Equilibrium play and best response to (stated) beliefs in normal form games. Games and Economic BehaviorDoi:10.1016/j.geb.2008.03.003.

Ross, G. L., House, P., 1977. False consensus effect: An egocentric bias in social perception and attribution processes. Journal of Experimental Social Psychology 13, 279 – 301.

Rubinstein, A., 2007. Instinctive and cognitive reasoning: Study of response times. The Economic Journal 117, 1243 – 1259.

Selten, R., Ockenfels, A., 1998. An experimental solidarity game. Journal of Economic Behavior and Organization 34, 517 – 539.

Stahl, D., Wilson, P., 1994. Experimental evidence on players' models of otherplayers. Journal of Economic Behavior and Organization 25, 309–327.

Stahl, D., Wilson, P., 1995. On players' models of other players: Theory and experimental evidence. Games and Economic Behavior, 10, 218–254.

Takeuchi, A., 2007. Jikken ni okeru senryakuteki suiron nouryoku no eikyou (the effect of strategic sophistication in experiments). In: Kohno, M., Saijo, T. (Eds.), Syakaikagaku no Jikken Apuro-chi. Keiso Shobo, Ch. 7, pp. 171 – 198.

Vyrastekova, J., Funaki, Y., Takeuchi, A., 2008. Strategic vs. non-strategic motivations of sanctioning, cent-ER Discussion Paper.

Wang, J. T., T., Spezio, M., Camerer, C. F., 2008. Pinocchio's pupil: Using eyetracking and pupil dilation to understand truth-telling and deception in sender-receiver games.

A Eye Tracking

The Eye-tracking data was recorded using The Tobii Eye Tracker 1750. The Tobii Eye Tracker uses binocular eye tracking, where the data are recorded from both eyes at the same time, which allows for more robust recording. The experiment was conducted using the dual setup of the computer, and the camera used for recording was mounted into The Tobii Eye Tracker, hidden in the black surface where it could not easily be recognized by the subjects. There were no head rests, chin rests or bite bars which restricts the subjects' head movements. The calibration was done using 5 points calibration after the understanding test of the STS Treatment.

The data was analyzed using the Clear View 2.7.0. The Fixation Filter was set with fixation radius of 30 pixel and minimum duration of 100ms. The data is also filtered by the validity codes set in the program. The gaze points used for analysis are the ones where the program could recognize and record both eyes, and all other gaze point data were excluded. Clear View also allows us to define the Area of Interest (AOI) on the screens used in the experiment. Clear View lists all the filtered gaze data in the AOI including the starting time of the fixation and the duration. The data we used for the analysis are the exported AOI data from the Clear View. In the analysis conducted in Section 4.2, the AOI was set to the outer edge of each matrix, right at the point where the color changes to white.

B Instruction for STS Treatment

Instruction[29]

This experiment consists of 14 rounds. In the beginning of the 14 rounds, you will be randomly assigned to one of the two groups. Which group you are assigned will not have an effect on your earnings.

In each round, you will be anonymously matched with one of the participants in the other group, a different one in each round. We will refer to the other participant as "s/he." In each round, you and s/he will be presented with a decision problem. Each of you, separately and independently, will make a CHOICE. Together, the two choices will determine the numbers of POINTS each of you earn that round. Points translate into your payment according to exchange rate

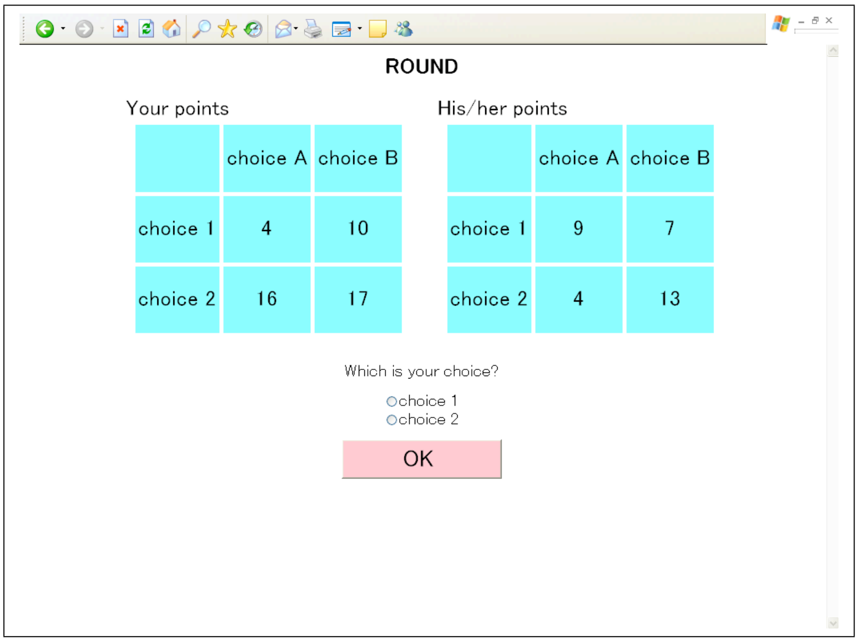
1 Point = 0.25 Euros.

Once a round is over, you will not be able to change your decision in that round. Neither you nor the other participants will learn the other's decisions or the points earned in any round until you come to pick up your earning. When you come and pick up the money, you will be able to see a list of your

[29] In the experiment analyzed in this paper, there was additional cover instruction. Also, there were some difference in the explanation of matching protocol in the sessions with four subjects each. The instruction used in the experiment are, along with the instruction for the Guess treatment, available as the supplementary material on the website.

decisions, the decision of the participant you were matched with, and the points you had earned in each round.

The next page displays an illustrative decision problem and its table of points. IT IS ONLY AN ILLUSTRATION, the decision problems you will face in the 14 rounds will be different from this one, and will change each round.



In the actual decision problems, you will be shown a table like this (but with different numbers of points) on your screen, and is asked to choose one of the possible choices, here labeled Choice 1 and Choice 2. The other participant with whom you are matched will be asked, independently, to choose one of his/her possible choices, here labeled Choice A and Choice B. The combination of your decision and his/her decision will determine the points you and s/he will earn in a round. Your points appear in the table on the left side of the screen, underneath the label ”Your Points”. His/her points appear in the table on the right side of the screen, underneath the label ”His/her Points”. To interpret the tables ”Your Points” and ”His/her Points”, consider the possible outcomes of your and his/her choices:

- If you choose Choice 1 and s/he chooses Choice A, you earn 4 points.
- If you choose Choice 1 and s/he chooses Choice A, s/he earns 9 points.
- If you choose Choice 1 and s/he chooses Choice B, you earn 10 points.
- If you choose Choice 1 and s/he chooses Choice B, s/he earns 7 points.
- If you choose Choice 2 and s/he chooses Choice A, you earn 16 points.
- If you choose Choice 2 and s/he chooses Choice A, s/he earns 4 points.
- If you choose Choice 2 and s/he chooses Choice B, you earn 17 points.
- If you choose Choice 2 and s/he chooses Choice B, s/he earns 13 points.

In each round of the actual decision problems, you will see new tables and you will be matched with a different participant. The points that you and s/he earn will depend on both your choice, and the choice of the other participant.

Please be sure that you understand the table. Ask the experimenter if you would like further explanation.

Different table formats

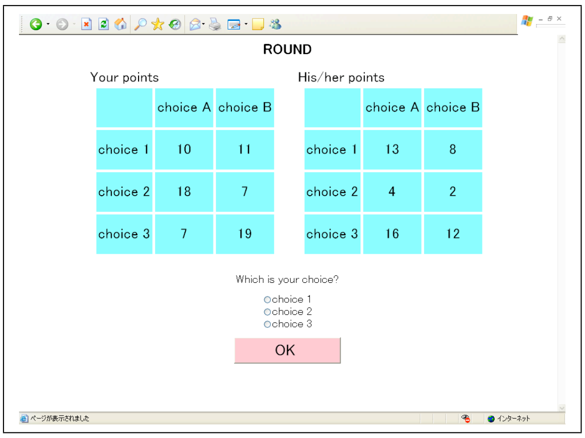
In some rounds of the experiment, you will be asked to choose one of THREE possible choices, labeled Choice 1, Choice 2, and Choice 3; while the other participant with whom you are matched will be asked to choose from TWO choices, as before. Thus, the table of points will have an extra row of boxes (see "Screen 1" at the bottom left).

In some other rounds, s/he will be asked to choose one of THREE possible choices, labeled Choice A, Choice B, and Choice C; while you will be asked to choose from TWO choices. Thus, the table of points will have an extra column of boxes (see "Screen 2" at the bottom right).

The Screen of the other participant

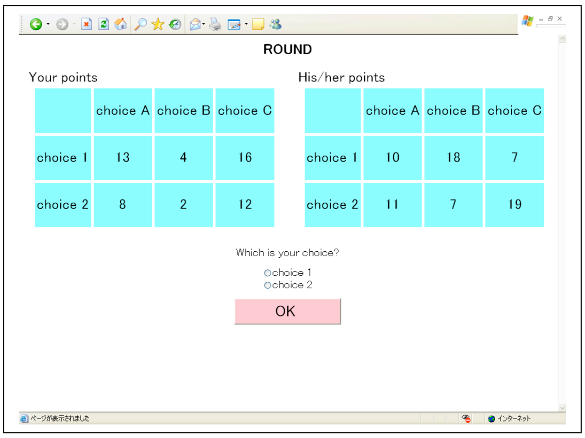
You may wonder whether you are going to be choosing your choice from numbered choices (like Choice 1 or Choice 2) or from alphabetized choices (like Choice A or Choice B). In the experiment, the points of the other participant will be rotated, so that every participant will be making their choice from numbered choices, like Choice 1 or Choice 2. The two example screens on this page are chosen to illustrate this. If you are faced with "Screen 1," the other participant whom you are matched with will be facing "Screen 2," and vice versa. If you carefully check the points on Screens 1 and 2, you will find the points on " Screen 2 " is rotated but the two are the same decision problem.

Screen 1



If you are faced with this screen

Screen 2



The Other Participant will be faced with this screen

Making decisions

To complete a given round, you must make a choice. Remember that the number of points you obtain in any given round will depend on both yours AND her/his choices. Your possible choices will be displayed below the table of points. To make your choice in this decision, click on one of your possible choices and then click OK.

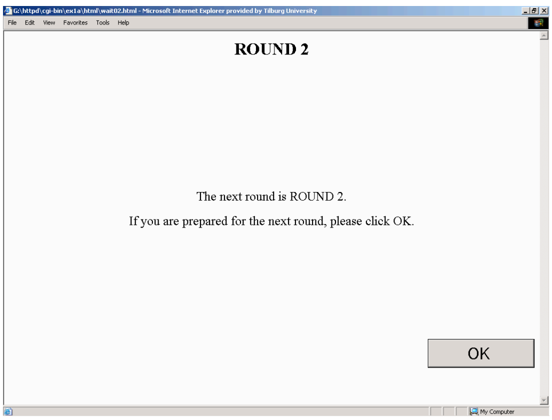
Make sure that you clicked on the correct choice before pressing OK. Once you have press OK, you will not be able to change your decision.

Matching protocol

In each round of the experiment, you will be anonymously matched with a different participant from the other group and both of you will face the same interdependent decision problem. Your decisions in a round will not influence the matching of the participants or the assignment of decision problems in later rounds. Your identity and the identities of the other participants will never be revealed. Each participant you are matched with will receive the same instructions as you and will face the same kind of screen display.

Other Screens

At the beginning of each round, you will see a screen like below, asking whether you are ready to proceed to the next round. If you wish, you may rest before proceeding. You are allowed to make your decisions at your own speed. However, we ask you not to rest during the screen with the Tables.



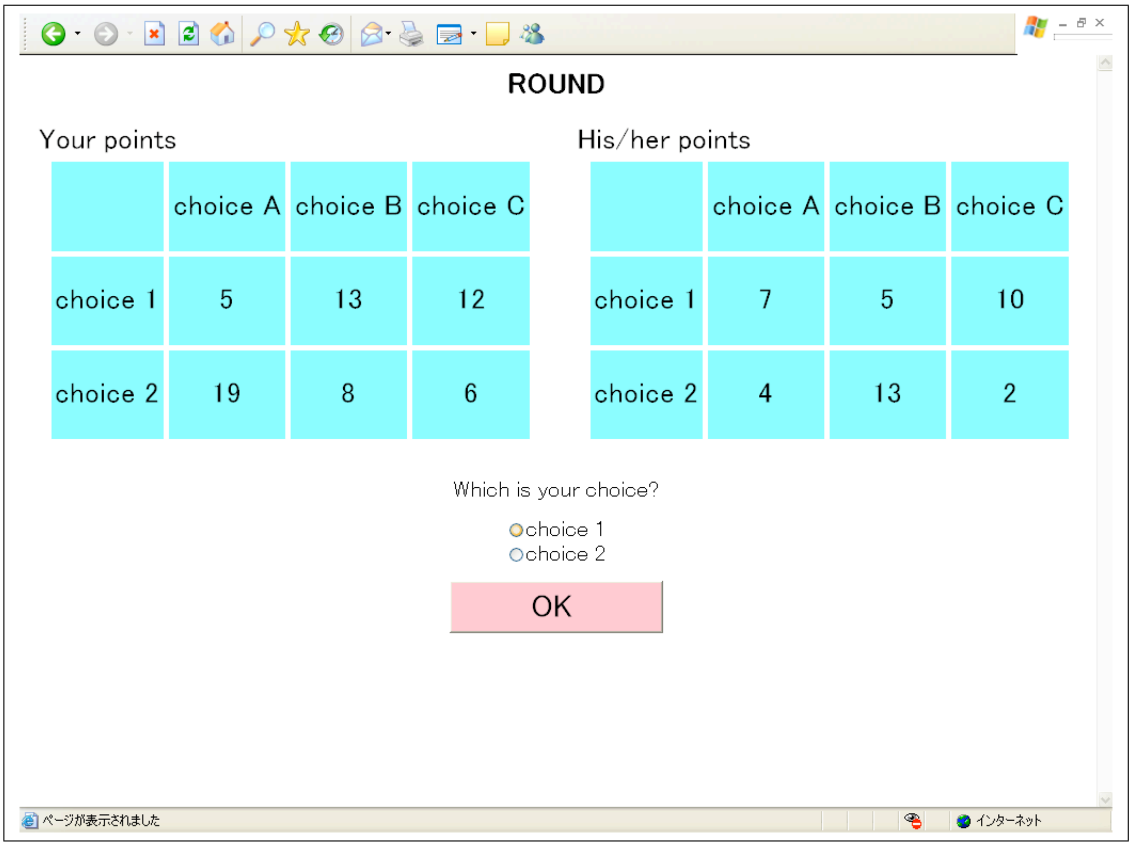
PAYMENT

After you have made your decisions for all 14 rounds, your payment will be determined according to the number of points you earned, as follows: When receiving the participation fee today, you will draw a card numbered 1 to 14. The number you draw will be the paid round. Outcome of this randomly selected round will be paid to you with the exchange rate 0.25 Euros per per point. You will be paid your earnings in cash, in private.

UNDERSTANDING TEST

You will now take a short UNDERSTANDING TEST. After you finish the TEST raise your hand and the experimenters will come and check your answers privately. If you have any questions during or after the test, you are always free to ask.

We do not have a trial round. Please make sure that you understand the procedure.



ROUND

Your points

	choice A	choice B	choice C
choice 1	5	13	12
choice 2	19	8	6

His/her points

	choice A	choice B	choice C
choice 1	7	5	10
choice 2	4	13	2

Which is your choice?

choice 1
choice 2

OK

ページが表示されました　インターネット

Answer the following questions using the table of points above. Note that in the table you have two choices and s/he has three.

1. If you choose Choice 1 and s/he chooses Choice A, how many points will you earn?
2. If you choose Choice 2 and s/he chooses Choice A, how many points will you earn?
3. If you choose Choice 2 and s/he chooses Choice B, how many points will s/he earn?
4. If you choose Choice 1 and s/he chooses Choice C, how many points will s/he earn?

YOU HAVE JUST COMPLETED THE TEST.

Please tell the experimenter that you are finished. We will come to correct your test.

C Calculation of the Probabilities in Table 2

$\bar{P}_{k,j}$ is the probability in which a type **H** player will not be classified as **H** due to small decision error. As we had explained, we consider the case where, when a player is a certain type, s/he will take the action of the corresponding type, but will make a mistake with probability of ε and chose each action randomly. So in games with t actions ($t = 2, 3$), players take the action of their true type with probability of $1-(t-1)\varepsilon/t$. $\bar{P}_{k,j}$ can be calculated by $1-\bar{p}_k^{MPM}(\varepsilon)\times\bar{p}_j^H(\varepsilon)$, where $\bar{p}_k^{MPM}(\varepsilon)$ is the probability of classifying the MPM type subject as MPM when k errors are allowed in games with dominant action and $\bar{p}_j^H(\varepsilon)$ is the probability of classifying **H** type subject as **H** when j errors are allowed in games solvable by 2 steps, given that s/he is always classified as MPM. Since there are 4 games with 2 actions and 1 games with 3 actions in games with dominant action and 5 games with 2 actions and 2 games with 3 actions in games solvable by 2 steps of iterated dominance, these two probabilities are calculated as follows. For $k = 0$, $\bar{p}_0^{MPM}(\varepsilon) = \big(1-\frac{\varepsilon}{2}\big)^4\big(1-\frac{2\varepsilon}{3}\big)$, for $k = 1, ..., 4$,

$$\bar{p}_k^{MPM}(\varepsilon) = \sum_{t=0}^{1} {}_4C_{k-t}\big(1-\frac{\varepsilon}{2}\big)^{4+t-k}\big(\frac{\varepsilon}{2}\big)^{k-t}\big(1-\frac{2\varepsilon}{3}\big)^{1-t}\big(\frac{2\varepsilon}{3}\big)^t + \sum_{l=0}^{k-1}\bar{p}_l^{MPM}(\varepsilon),$$

and 1 for $k = 5$. For $j = 0$, $\bar{p}_0^H(\varepsilon) = \big(1-\frac{\varepsilon}{2}\big)^5\big(1-\frac{2\varepsilon}{3}\big)^2$, for $j = 1, ..., 6$,

$$\bar{p}_j^H(\varepsilon) = \sum_{t=0}^{2} {}_5C_{j-t}\times {}_2C_t\big(1-\frac{\varepsilon}{2}\big)^{5+t-j}\big(\frac{\varepsilon}{2}\big)^{j-t}\big(1-\frac{2\varepsilon}{3}\big)^{2-t}\big(\frac{2\varepsilon}{3}\big)^t + \sum_{l=0}^{j-1}\bar{p}_l^H(\varepsilon)$$

and 1 for $j = 7$.

Next, $P_{k,j}$ is the probability in which a random player who chose each action with a same probability will be classified as **H**. It can be calculated by multiplying the probability of classifying a random player into MPM when k errors are allowed for games solvable by 1 step (let us denote this probabilities as p_k^{MPM}) and the probability of classifying a random player into **H** given that the random player would always be classified as MPM and j errors are allowed in games solvable by 2 steps of iterated dominance (let us denote this probability as p_j^{H}). So, $P_{k,j} = p_k^{MPM}\times p_j^H$.

p_k^{MPM} can be calculated as follows. For k = 0, $p_0^{\text{MPM}} = \big(\frac{1}{2}\big)^4\big(\frac{1}{3}\big)$,

$$p_k^{\text{MPM}} = \sum_{t=0}^{1} {}_4C_{k-t}\times\big(\frac{1}{2}\big)^4\big(\frac{1}{3}\big)^{1-t}\big(\frac{2}{3}\big)^t + \sum_{l=0}^{k-1} p_l^{\text{MPM}} \text{ for } k = 1, ..., 4$$

and 1 for $k = 5$. Similarly, p_j^{H} can be calculated as follows. For j = 0, $p_0^{\mathrm{MPM}} = \left(\frac{1}{2}\right)^5 \left(\frac{1}{3}\right)^2$,

$$p_j^{\mathrm{H}} = \sum_{t=0}^{2} {}_5C_{j-t} \times {}_2C_t \Big(\frac{1}{2}\Big)^5 \Big(\frac{1}{3}\Big)^{2-t} \Big(\frac{2}{3}\Big)^t + \sum_{l=0}^{j-1} p_l^{\mathrm{H}} \ \text{ for } \ j = 1, ..., 6$$

and 1 for $j = 7$.