




SPECIAL ISSUE ARTICLE

# Measuring higher-order rationality with belief control

Wei James Chen<sup>1</sup> , Meng-Jhang Fong<sup>2</sup>  and Po-Hsuan Lin<sup>3</sup> 

<sup>1</sup>Department of Agricultural Economics, National Taiwan University, Taipei, Taiwan

<sup>2</sup>California Institute of Technology, Pasadena, CA, USA

<sup>3</sup>Department of Economics, University of Virginia, Charlottesville, VA, USA

**Corresponding author:** Meng-Jhang Fong; Email: [mengjhangfong@gmail.com](mailto:mengjhangfong@gmail.com)

(Received 14 September 2023; revised 15 February 2025; accepted 16 February 2025)

## Abstract

Determining an individual's strategic reasoning capability based solely on choice data is a complex task. This complexity arises because sophisticated players might have non-equilibrium beliefs about others, leading to non-equilibrium actions. In our study, we pair human participants with computer players known to be fully rational. This use of robot players allows us to disentangle limited reasoning capacity from belief formation and social biases. Our results show that, when paired with robots, subjects consistently demonstrate higher levels of rationality, compared to when paired with human players. Furthermore, players' rationality levels are relatively stable across games when paired with robot players, even though those with intermediate rationality levels exhibit inconsistency across games. Leveraging our experimental design, we identify and document potential causes of this inconsistency.

**Keywords:** Guessing game; higher-order rationality; level- $k$ ; ring game

**JEL Codes:** C72; C92; D83; D90

## 1. Introduction

Understanding whether individuals make optimal choices in strategic environments is a fundamental question in economics. Unlike individual decision-making, a game involves multiple players whose payoffs depend on each other's choice. In this setting, achieving equilibrium requires a player to exhibit both first-order rationality and higher-order rationality. This necessitates that players are not merely rational themselves but also operate under the assumption that their counterparts are rational. Furthermore, they must believe that other participants consider them to be rational, and this belief cascades infinitely. As a result, in equilibrium, each player's assumptions about the strategies of their peers match the actual strategies employed, allowing them to optimally respond.

However, expecting players to engage in iterative reasoning and demonstrate infinite levels of rationality is notably demanding, especially when viewed empirically. This is evidenced by well-documented instances of players diverging from equilibrium play (see, for example, Camerer 2003). Given these empirical discrepancies, a significant volume of research has been dedicated to determining the extent of iterative reasoning an individual can realistically execute within different contexts.

Apart from exploring the extent of iterative reasoning an individual can undertake, this paper delves into another crucial, related query: Is there consistency in an individual's depth of strategic reasoning across various games? Measuring strategic reasoning abilities of interacting individuals

can facilitate our understanding and predictions of individuals' behavioral patterns. It also helps us evaluate whether the observed non-equilibrium actions are driven by bounded rationality or by other factors. Nevertheless, if we observe no regularity when measuring one's depth of strategic reasoning in different environments, there may not even exist such a persistent trait called "strategic thinking ability."

The main challenge behind inferring individual strategic reasoning ability from choice data is that the strategic sophistication revealed by one's choices does not directly imply the maximum steps of iterative reasoning one is able to perform. As noted by Ohtsubo & Rapoport (2006),<sup>1</sup> a player's observed depth of reasoning is determined not only by their reasoning capability but also by their beliefs about the opponents' (revealed) sophistication, a notion supported by empirical evidence in Agranov et al. 2012 and Alaoui and Penta (2016). An individual who can carry out more than  $k$  steps of reasoning would act as a  $k$ th-order rational player when they believe that their opponent exhibits  $(k - 1)$ th-order rationality. In other words, measuring an individual's revealed strategic sophistication only yields a lower-bound estimate of their actual sophistication. In addition, psychological factors other than bounded rationality such as lying aversion and fairness concern may also motivate a player to deviate from an equilibrium (Cooper & Kagel, 2016). Without controlling for a player's beliefs and social preferences, the estimation of their strategic reasoning ability could be unstable and lack external validity.

In a study on bounded strategic sophistication by Georganas et al. (2015), a question similar to the one posed in this paper is explored. In their research, participants play two distinct families of games. Although their study does not extensively control for participants' beliefs, it reveals a limited persistence of individual strategic sophistication between the two families of games.<sup>2</sup>

In this paper, we demonstrate a method to test the stability of individual strategic sophistication and to possibly pin down the upper bound of an individual's depth of strategic reasoning in the lab: having human subjects interact with equilibrium-type *computer* players induced by infinite order of rationality. By informing human players that they are facing fully rational computer players, we are able to unify players' expectations about their opponents. Additionally, introducing computer players precludes the possible effect of social preferences (Houser & Kurzban, 2002; Johnson et al., 2002; Van den Bos et al., 2008). Thus, human players with an infinite order of rationality are expected to select an equilibrium strategy. In this setting, out-of-equilibrium actions would provide us with a solid ground to identify an individual's order of rationality for inferring their strategic reasoning ability since those actions are likely driven by bounded rationality.

To investigate the stability of individual strategic sophistication across games, we conduct an experiment with two classes of dominance-solvable games, ring games and guessing games. Proposed by Kneeland (2015) for identifying higher-order rationality, an  $n$ -player ring game can be characterized by  $n$  payoff matrices and has the following ring structure: the  $k$ th player's payoff is determined by the  $k$ th player's and  $(k + 1)$ th player's actions, and the payoff of the last ( $n$ th) player, who has a strictly dominant strategy, is determined by the last and the first player's actions. We employ guessing games that represent a symmetric variant of the two-person guessing games studied by Costa-Gomes and Crawford (2006), in which a player's payoff is single-peaked and maximized if the player's guess equals its opponent's guess multiplied by a predetermined number.<sup>3</sup>

<sup>1</sup>"Subjects who go through several levels of reasoning and figure out the equilibrium solution to the game, will in general not invoke the maximum depth of reasoning precisely because they do not assume—and perhaps should not assume—that the other  $n - 1$  players are as smart as they are" (Ohtsubo & Rapoport, 2006, p. 45).

<sup>2</sup>Another study that reports inconsistent depth of reasoning across games is Cooper et al. (2024), which examines the comparative statics predictions of the level- $k$  model without controlling for participants' beliefs. Note that the idea of examining cross-game persistence of reasoning depth can be traced back to Stahl and Wilson (1995), who found that 72% of subjects had a stable depth of reasoning, though they focused on a *single* family of games.

<sup>3</sup>The guessing game we implement in this paper diverges from the standard beauty contest game, primarily because the standard beauty contest game is not strictly dominant solvable. However, it is worth noting that if the beauty

Among the games that have been used to study strategic reasoning, we choose to implement ring games and guessing games in our experiment for two reasons. First, our instruction of a fully rational computer player's behavior is tailored to align with the payoff structure of dominance-solvable games, in which the computer players' actions can be unambiguously determined (see [Section 5.1](#) for details). Furthermore, these dominance-solvable games enable a structure-free identification approach, leveraging the notion of rationalizable strategy sets (Bernheim, 1984; Pearce, 1984). The core idea behind this identification approach is that, within a dominance solvable game, we can gauge an individual's depth of reasoning by assessing how many rounds of iterated deletion of dominated strategies the individual's chosen action would survive. Importantly, this approach does not impose structural assumptions on (the beliefs about) non-rational players' behavior. Therefore, these classes of games provide a plausible, structure-free method to empirically categorize individuals into distinct levels of rationality.

Second, we intend to implement two types of games that are sufficiently different so that, if we observe any stability in individual strategic reasoning levels across games, the stability does not result from the similarity between games. We believe that ring games and guessing games are dissimilar to each other. On the one hand, a ring game is a four-player discrete game presented in matrix forms. On the other hand, a guessing game is a two-player game with a large strategy space, which is more like a continuous game. In fact, Cerigioni et al. 2019 report that the correlation of their experimental subjects' reasoning levels between ring games and beauty contest games is only 0.10. Although not intended to provide conclusive evidence from a limited number of games, we believe our study takes an important step toward investigating the consistency of reasoning levels across diverse game types, in line with recent literature encouraging further examination of cross-game stability.

Our experiment comprises two treatments within each game family: the Robot Treatment and the History Treatment. In the Robot Treatment, subjects encounter computer players employing equilibrium strategies. In the History Treatment, subjects confront choice data from human players in the Robot Treatment. The History Treatment simulates an environment where human subjects interact without displaying social preferences and serves two main objectives. First, by examining if a subject's observed rationality level in the Robot Treatment exceeds that in the History Treatment, we can evaluate whether the subject responds to equilibrium-type computer players by employing a strategy that reaches their full capacity for strategic reasoning. Second, by comparing the individual orders of rationality inferred from data in both the Robot and History Treatments, we can investigate whether the introduction of robot players contributes to stabilizing observed rationality levels across various games.

Overall, our findings indicate that strategic reasoning ability may be a persistent personality trait deducible from choice data when subjects interact with robot players in strategic scenarios. Relative to interactions with human opponents, we observe a larger proportion of participants adopting equilibrium strategies and demonstrating higher levels of rationality. This observation is supported by both our between- and within-subject statistical analyses, underscoring the effectiveness of our Robot Treatment and implying that the rationality levels exhibited in this treatment potentially approach subjects' strategic thinking capacity.<sup>4</sup>

Furthermore, our investigation reveals that subjects' rationality levels remain relatively stable across distinct game classes when interacting with robot players. In terms of absolute levels, a substantial number of first-order and fourth-order rational players retain their respective types while transitioning from ring games to guessing games. In the Robot Treatment, approximately 38% of

---

contest game involves only two players, then it becomes dominant solvable (Chen & Krajbich, 2017; Grosskopf & Nagel, 2008).

<sup>4</sup>One might doubt if a subject has the motivation to act rationally upon the presence of an opponent with a (much) higher rationality level than the subject has. In [Section 7.1](#), we argue that a subject does have the incentive to exhibit the highest order of rationality they can achieve when they know their opponent is at least as rational as themselves.

subjects exhibit constant rationality levels across games.<sup>5</sup> Statistical tests demonstrate that this stability in rationality levels, unlike in the History Treatment, cannot be attributed to independent type distributions, highlighting the impact of our belief manipulation regarding opponents' reasoning depth. Despite the relatively stable distribution of rationality levels, players with intermediate rationality (second-order and third-order rationality) display inconsistency across games. To address this, we introduce a diagnostic classification, which demonstrates that these players are more inclined to avoid dominated strategies rather than exhibit consistent levels of rationality, particularly when uncertain about others' strategies (as seen in the History Treatment).

A subject's performance in other cognitive tests could potentially hold predictive power regarding their strategic reasoning performance in games. As such, we incorporate tasks measuring cognitive reflection, short-term memory, and backward induction abilities (see [Section 5.3](#) for details) into our experiment. We observe that a subject's cognitive reflection and backward induction abilities are positively correlated with their levels of rationality, whereas no significant correlation is found with their short-term memory capacity.

The rest of the paper proceeds as follows. The next subsection reviews the related literature. [Section 2](#) summarizes the theoretical framework upon which our identification approach and hypotheses to be tested are based. [Section 3](#) describes the ring games and guessing games implemented in our experiment. [Section 4](#) discusses how we identify a subject's rationality level given choice data. [Section 5](#) presents our experimental design and the hypotheses to be tested. The experimental results are reported in [Section 6](#). In [Section 7](#), we discuss the validity and limitations of our belief control approach for the Robot Treatment, as well as the results of the diagnostic classification of types. Finally, [Section 8](#) concludes. The complete instructions of our experiment can be found in Supplementary Information.<sup>6</sup>

### 1.1. Related literature

Over the past thirty years, various researchers have theoretically studied the idea of limited depth of reasoning, including Selten (1991); Selten (1998), Aumann 1992, Stahl (1993), Camerer et al. (2004), Alaoui and Penta (2016); Alaoui and Penta (2022), Lin 2023, and Lin and Palfrey (2024). In addition to theoretical contributions, Nagel (1995) conducts the first experiment on beauty contest games and introduces the level- $k$  model to describe non-equilibrium behavior. The behavior that can be explained by assuming level- $k$  reasoning has subsequently been observed in a variety of game types, including matrix games (e.g., Costa-Gomes et al., 2001; Crawford & Iriberri, 2007a; Stahl & Wilson, 1994; Stahl & Wilson, 1995), investment games (e.g., Rapoport & Amaldoss, 2000), guessing games (e.g., Costa-Gomes & Crawford, 2006), undercutting games (e.g., Arad & Rubinstein, 2012), auctions (e.g., Crawford & Iriberri, 2007b), and sender-receiver games (e.g., Cai & Wang, 2006; Fong & Wang, 2023; Wang et al., 2010), though the list is not exhaustive.

Unlike the literature that primarily investigates individuals' strategic sophistication within the context of a single specific game, our work, which is closely related to Georganas et al. (2015) (hereinafter, GHW), centers on the examination of the consistency of strategic sophistication across different games. In particular, we follow the language of GHW to formalize our hypotheses to be tested.<sup>7</sup> Although both GHW and this paper experimentally investigate whether a subject's sophistication type persists across games, our study differs from GHW in several ways. First, we substitute the ring games for the undercutting games in GHW and use a simplified, symmetric version of the guessing games. Second, we employ an identification strategy distinct from the standard level- $k$  model to

<sup>5</sup>The constant rationality level hypothesis is the strictest requirement for the stability of rationality levels across games. In Online Appendix B, we explore two weaker notions of stability and find that players' rationality levels are more stable across games in the Robot Treatment, even when considering these weaker notions of stability.

<sup>6</sup>The provided instructions are originally in Chinese and have been translated into English.

<sup>7</sup>For a brief summary of the model in GHW, see [Section 2.1](#); also, see [Section 5.2](#) for the hypotheses.

determine a subject's strategic sophistication. We use dominance solvable games in order to identify higher-order rationality without imposing strong and ad hoc assumptions on players' first-order beliefs, which can in turn reduce the noise in the estimation of individual reasoning depth using a level- $k$  model.<sup>8</sup> More importantly, we control for human subjects' beliefs about opponents' sophistication (and social preferences) using computer players. As a result, we observe a higher correlation in subjects' types across games compared to GHW, in which subjects are matched with each other.

Ring games, first utilized for identifying higher-order rationality by Kneeland (2015), are subsequently studied by Lim and Xiong (2016) and Cerigioni et al. (2019), who investigate two variants of the ring games. In this study, we follow the *revealed rationality approach* adopted by Lim and Xiong (2016) and Cerigioni et al. (2019) as our identification approach (discussed in Section 4). It is worth noting that Cerigioni et al. (2019) also find little correlation in subjects' estimated types across various games, including ring games, e-ring games,  $p$ -beauty contests, and a  $4 \times 4$  matrix game. Again, our results suggest that the lack of persistence in the identified order of rationality at the individual level is driven by subjects' heterogeneous beliefs about the rationality of their opponents.

Indeed, several empirical studies have shown that beliefs about others' cognitive capacity for strategic thinking can alter a player's strategy formation. Friedenberg et al. (2018) indicate that some non-equilibrium players observed in the ring games (Kneeland, 2015) may actually possess high cognitive abilities but follow an irrational behavioral model to reason about others. Alternatively, Agranov et al. (2012), Alaoui and Penta (2016), Gill and Prowse (2016), and Fe et al. (2022) find that, in their experiments, subjects' strategic behavior is responsive to the information they receive about their opponents' strategic abilities.<sup>9</sup> The designs of experiments allow them to manipulate subjects' beliefs, whereas we aim to elicit and identify individual strategic capability by unifying subjects' beliefs about opponents.

Some recent studies have tried to distinguish between non-equilibrium players who are limited by their reasoning abilities and players who are driven by beliefs. Identifying the existence of ability-bounded players is important since, if non-equilibrium behavior is purely driven by beliefs, it would be unnecessary to measure an individual's reasoning depth. Jin (2021) utilizes a sequential version of ring games, finding that around half of the second-order and third-order rational players are bounded by ability. Alaoui et al. (2020) also report the presence of ability-bounded subjects by showing that an elaboration on the equilibrium strategy shifts the subjects' level- $k$  types toward higher levels. Overall, the existence of both ability-bounded and belief-driven players in the real world indicates the need for an approach that can measure individual reasoning ability without the impact of beliefs. Whereas Jin (2021) and Alaoui et al. (2020) do not pin down the belief-driven players' actual ability limit, we aim to directly measure each subject's strategic ability.

Bosch-Rosa and Meissner (2020) propose an approach to test a subject's reasoning level in a given game: letting a subject play against herself (i.e., an "one-person" game). Specifically, in their study, each subject acts as both players in a modified two-person  $p$ -beauty contest (Chen & Krajbich, 2017; Grosskopf & Nagel, 2008), in which a player's payoff decreases in the distance between their own guess and the average guess multiplied by  $p$ , and the subject receives the sum of the two players' payoffs.<sup>10</sup> The one-person game approach eliminates the impact of beliefs that arises from interacting with human players. However, a limitation of this approach is that it can only be applied to the game

<sup>8</sup>Burchardi and Penczynski (2014) conduct an experiment in a standard beauty contest with belief elicitation, finding heterogeneity in both level-0 beliefs and level-0 actions within a game.

<sup>9</sup>In Agranov et al. (2012), subjects play against each other, graduate students from NYU Economics Department, or players taking uniformly random actions. In Alaoui and Penta (2016), subjects play against opponents majoring in humanities, majoring in math and sciences, getting a relatively high score, or getting a low score in a comprehension test. In Gill and Prowse (2016) and Fe et al. (2022), subjects play against opponents with similar or differing performance in cognitive tests.

<sup>10</sup>Bosch-Rosa and Meissner (2020) report that 69% of the subjects do not select the equilibrium action (0, 0) when playing the one-person game, which echoes the findings of the presence of ability-bounded players in Jin (2021) and Alaoui et al. (2020).

in which the equilibrium is Pareto optimal. For instance, it would be rational for a payoff-maximizing subject to deviate from the equilibrium and choose (Cooperate, Cooperate) in the prisoner's dilemma since (Cooperate, Cooperate) maximizes the total payoff of both players even though those are not equilibrium strategies.<sup>11</sup> In this study, we employ an alternative approach that overcomes this limitation to measure rationality levels: letting a subject play against equilibrium-type computer players (i.e., the Robot Treatment).

Similar to the motivation of our Robot Treatment, Devetag and Warglien (2003), Grehl and Tutić (2015), and Bayer and Renou (2016) also employ rational computer players to mitigate the impact of beliefs and social preferences on individual decisions in their experiments. Grehl and Tutić (2015) and Bayer and Renou (2016) explore players' ability to reason logically about others' types in the incomplete information game known as the dirty faces game. In contrast, our study focuses on investigating whether playing against computers can provide a robust measure of strategic reasoning ability across different families of games with complete information. Additionally, Devetag and Warglien (2003) examine the relationship between short-term memory performance and conformity to standard theoretical predictions in strategic behavior, finding a positive correlation between the two. Building on this, we include a memory task to investigate whether the lack of significant predictive power of short-term memory on reasoning levels observed in GHW is influenced by uncontrolled beliefs, and to offer a robustness check for the findings of Devetag and Warglien (2003) in different settings.

In previous studies on strategic reasoning, equilibrium-type computer players have been introduced into laboratory experiments to induce human players' equilibrium behavior (e.g., Costa-Gomes & Crawford, 2006; Meijering et al., 2012) and to eliminate strategic uncertainty (e.g., Hanaki et al., 2016).<sup>12</sup> In contrast, our aim is to utilize computer players to uncover individual strategic reasoning ability. Our study contributes to the literature by demonstrating that introducing robot players can induce human subjects to exhibit stable reasoning levels across games, thus providing a solid foundation for measuring individual strategic ability.

## 2. Theoretical framework

### 2.1. The model in GHW

To formalize the idea of the depth of rationality and the hypotheses we are going to test, we introduce the model and notations used in GHW. In their model, an  $n$ -person normal form game  $\gamma \in \Gamma$  is represented by  $(N, S, \{u_i\}_{i \in N})$ , where  $N = \{1, \dots, n\}$  denotes the set of players,  $S = S_1 \times \dots \times S_n = \prod_{i=1}^n S_i$  denotes the strategy sets, and  $u_i : S \rightarrow \mathbb{R}$  for  $i \in N$  denotes the payoff functions.

Player  $i$ 's strategic ability is modeled by two functions  $(c_i, k_i)$ . Let  $T$  be the set of *environmental parameters*, which captures the information a player observes about their opponents' cognitive abilities. The function  $c_i : \Gamma \rightarrow \mathbb{N}_0$  represents  $i$ 's *capacity* for game  $\gamma$ , and the function  $k_i : \Gamma \times T \rightarrow \mathbb{N}_0$  represents  $i$ 's (realized) *level* for game  $\gamma$ . A player's level for a game is bounded by their capacity, so  $k_i(\gamma, \tau_i) \leq c_i(\gamma)$  for all  $\gamma, \tau_i \in T$ , and  $i \in N$ . The goal of our experiment is to measure  $c_i(\gamma)$  and to test if  $c_i(\gamma)$  (or  $k_i(\gamma, \tau_i)$ , after controlling for  $\tau_i$ ) exhibits any stability across different games (see Section 5.2 for further discussion).

### 2.2. Higher-order rationality

To characterize a player's behavior in the games, we define  $k$ th-order rationality (Bernheim, 1984; Lim & Xiong, 2016; Pearce, 1984) in the following way. Let  $R_i^k(\gamma)$  be the set of strategies that survive  $k$  rounds of iterated elimination of strictly dominated strategies (IEDS) for player  $i$ . Namely, a strategy  $s_i$  is in  $R_i^1(\gamma)$  if  $s_i$  is a best response to some arbitrary  $s_{-i}$ , and  $s_i$  is in  $R_i^{k'}(\gamma)$  if  $s_i$  is a best response to

<sup>11</sup> Also note that in the ring game G1, both the equilibrium strategy profile (P1:  $b$ , P2:  $c$ , P3:  $c$ , P4:  $b$ ) and a non-equilibrium strategy profile (P1:  $a$ , P2:  $b$ , P3:  $a$ , P4:  $a$ ) lead to a total payoff of 66 (see Figure 1).

<sup>12</sup> For a survey of economics experiments with computer players, see March (2021).



some  $s_{-i} \in R_{-i}^{k'-1}(\gamma)$  for  $k' > 1$ . We say that a player  $i$  exhibits *kth-order rationality* in  $\gamma$  if and only if  $i$  always plays a strategy in  $R_i^k(\gamma)$ . Note that given any game  $\gamma \in \Gamma$ ,  $R_i^{k+1}(\gamma) \subset R_i^k(\gamma)$  for all  $k \in \mathbb{N}_0$ . In other words, a player exhibiting *kth-order rationality* also exhibits *jth-order rationality* for all  $j \leq k$ .

We characterize boundedly rational behavior using higher-order rationality rather than the standard level- $k$  model employed by GHW to avoid the ad hoc assumption regarding level-0 players.<sup>13</sup> In the standard level- $k$  model, pinning down a level- $k$  player's strategy requires an assumption about the level-0 strategy. However, studies have reported variations in level-0 actions and level-0 beliefs across individuals (Burchardi & Penczynski, 2014; Chen et al., 2018). Consequently, an individual's identified level of reasoning can be sensitive to the structural assumptions of the level- $k$  model. Alternatively, the higher-order rationality approach avoids structural assumptions about (beliefs regarding) non-rational players' behavior, thereby providing a structure-free method for empirically categorizing individuals into distinct reasoning levels.

### 3. The games

We study two classes of games: the four-player ring games used in Kneeland (2015) for identifying individuals' higher-order rationality and a variant of the two-person guessing games first studied by Costa-Gomes & Crawford (2006) and used in GHW for identifying players' level- $k$  types.

#### 3.1. Ring games

A four-player ring game is a simultaneous game characterized by four  $3 \times 3$  payoff matrices. Figure 1 summarizes the structures of the two ring games, G1 and G2, used in our experiment. As shown in Figure 1, each player  $i \in \{1, 2, 3, 4\}$  simultaneously chooses an action  $a_i \in \{a, b, c\}$ . Player 4 and Player 1's choices determine Player 4's payoff, and Player  $k$  and Player  $(k + 1)$ 's choices determine Player  $k$ 's payoff for  $k \in \{1, 2, 3\}$ .

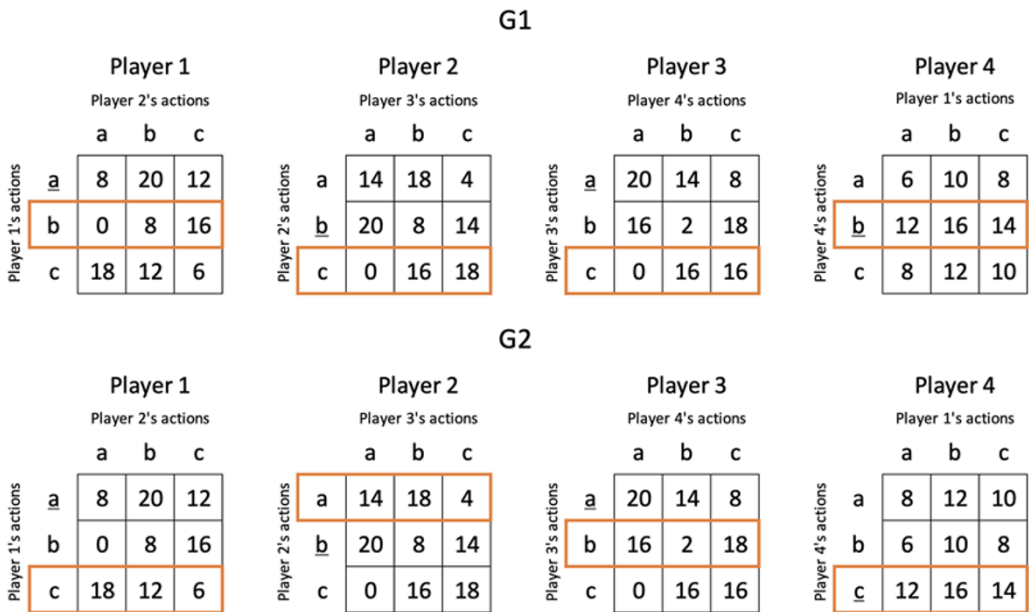
The payoff matrices for Player 1, 2, and 3 are identical in G1 and G2. However, the matrix for Player 4 differs between G1 and G2, with the rows corresponding to Player 4's actions ( $a, b, c$ ) interchanged, leading to different best replies in the subsequent matrices.

Specifically, Player 4 has a strictly dominant strategy in each ring game:  $b$  in G1 and  $c$  in G2. Given the payoff structure, a (first-order) rational individual will always choose  $b$  in G1 and  $c$  in G2 when acting as Player 4. By eliminating dominated strategies, an individual exhibiting second-order rationality will always choose  $c$  in G1 and  $b$  in G2 when acting as Player 3. Continuing this process iteratively, an individual exhibiting third-order rationality will always choose  $c$  in G1 and  $a$  in G2 when acting as Player 2, and an individual exhibiting fourth-order rationality will always choose  $b$  in G1 and  $c$  in G2 when acting as Player 1. Thus, the unique Nash equilibrium of G1 is Player 1, 2, 3, and 4 choosing  $b, c, c$ , and  $b$ , respectively, and for G2, Player 1, 2, 3, and 4 choosing  $c, a, b$ , and  $c$ , as highlighted in Figure 1.

Note that the payoff structures in our ring games are identical to those in Kneeland (2015), except that rows  $a$  and  $b$  are swapped for Player 4 in G1. This modification ensures that our equilibrium-predicted actions do not coincide with the *secure* actions (or max-min actions) in both G1 and G2, which maximize the total payoff sum over the opponents' possible actions, potentially encouraging subjects to choose the equilibrium strategy based on non-payoff-maximizing motives.<sup>14</sup>

<sup>13</sup>Let  $\nu : \mathbb{N}_0 \rightarrow \Delta(\mathbb{N}_0)$  be a player's belief about their opponents' levels. In the standard level- $k$  model,  $\nu(m) = \mathbb{1}\{m = 1\}$  for all  $m \geq 1$ , and a level-0 player  $i$ 's strategy is exogenously given as  $\sigma_i^0 \in \Delta(S_i)$ . For all  $s'_i \in S_i$ ,  $\sigma_i^k$  satisfies  $u_i(\sigma_i^k, \sigma_{-i}^{\nu(k)}) \geq u_i(s'_i, \sigma_{-i}^{\nu(k)})$ , where  $\sigma_{-i}^{\nu(k)} = (\sigma_1^{k-1}, \dots, \sigma_{i-1}^{k-1}, \sigma_{i+1}^{k-1}, \dots, \sigma_n^{k-1})$ . Here,  $u_i(\sigma)$  refers to  $E_\sigma[u_i(\sigma)]$ , where  $\sigma = (\sigma_1, \dots, \sigma_n)$  is a profile of mixed strategies (i.e.,  $\sigma_i \in \Delta(S_i)$ ). Notice that an individual exhibits *kth-order rationality* if and only if there exists some  $\sigma_{-i}^0$  such that the individual can be classified as a level- $k$  player under the standard level- $k$  model.

<sup>14</sup>A consequence of this modification is that the minimum possible payoff for the equilibrium strategy in G1 becomes 0 for Player 1, 2, and 3.



**Figure 1.** The ring games. The Nash equilibrium is highlighted with colored borders, and the secure actions are underscored

Adopting the same payoff structure as Kneeland’s design facilitates comparability between our results and hers.

3.2. Guessing games

In our experiment, the guessing game is a simultaneous two-player game parameterized by a constant  $p \in (0, 1)$ . We use  $p = 1/3, 1/2$  and  $2/3$  in our experiment. Each player  $i$  simultaneously chooses a positive integer  $s_i$  between 1 and 100. Player  $i$ ’s payoff strictly decreases in the difference between the number chosen by  $i$ ,  $s_i$ , and the number chosen by  $i$ ’s opponent multiplied by a constant  $p$ ,  $ps_{-i}$ . Specifically, player  $i$ ’s payoff is equal to  $0.2 \times (100 - |s_i - ps_{-i}|)$ . Thus, a payoff-maximizing player’s objective is to make a guess that matches their opponent’s guess times  $p$ . Note that, given  $p < 1$ , any action (integer) greater than or equal to  $\lfloor 100p + 0.5 \rfloor + 1$  is strictly dominated by  $\lfloor 100p + 0.5 \rfloor$  since  $|\lfloor 100p + 0.5 \rfloor - ps_{-i}| < |s'_i - ps_{-i}|$  for all  $s_{-i} \in \{1, \dots, 100\}$  and  $s'_i \in \{\lfloor 100p + 0.5 \rfloor + 1, \dots, 100\}$ .<sup>15</sup>

Given the payoff logic above, a (first-order) rational individual will choose an integer between 1 and  $K_1 \equiv \lfloor 100p + 0.5 \rfloor$ , since any action greater than  $K_1$  is strictly dominated by  $K_1$ . A second-order rational individual will believe the other player is first-order rational and choose a positive integer between 1 and  $\lfloor K_1p + 0.5 \rfloor$ , and so on. The unique equilibrium of the two-person guessing game is thus both players choosing 1.

4. Identification

Our model does not allow us to directly identify one’s higher-order rationality from choice data. For example, an equilibrium player will choose 1 in the guessing game with  $p = 1/2$ , while a player choosing 1 may have only performed one step of reasoning if their first-order belief is that their opponent guesses 2. Thus, observing a player  $i$  choosing a strategy in  $R^k_i(\cdot)$  for  $k > 1$  (in a

<sup>15</sup>For instance, in a guessing game with  $p = 1/3$ , every integer between 34 and 100 is dominated by 33; when  $p = 1/2$ , every integer between 51 and 100 is dominated by 50; when  $p = 2/3$ , every integer between 68 and 100 is dominated by 67.



**Table 1.** Predicted actions in the ring games under the revealed rationality approach

Level	Ring Games							
	P1		P2		P3		P4	
	G1	G2	G1	G2	G1	G2	G1	G2
R0	N/A		N/A		N/A		not (b, c)	
R1	N/A		N/A		not (c, b)		(b, c)	
R2	N/A		not (c, a)		(c, b)		(b, c)	
R3	not (b, c)		(c, a)		(c, b)		(b, c)	
R4	(b, c)		(c, a)		(c, b)		(b, c)	

finite number of rounds) does not imply that  $i$  exhibits  $k$ th-order rationality, which renders an individual's higher-order rationality unidentifiable. In fact, following the definition of  $R_i^k(\cdot)$ , we have  $R_i^{k+1}(\cdot) \subset R_i^k(\cdot)$  for all  $k \in \mathbb{N}_0$ . Namely, every strategy (except for the dominated actions) can be rationalized by some first-order belief.

Following the rationale of higher-order rationality, we use the *revealed rationality approach* (Lim & Xiong, 2016; Brandenburger et al., 2019; Cerigioni et al., 2019) as our identification strategy. As explained below, this approach allows us to identify individual higher-order rationality in a dominance-solvable game. Under the revealed rationality approach, we say that a player  $i$  exhibits  $k$ th-order revealed rationality if (and only if) we observe the player actually playing a strategy that can survive  $k$  rounds of IEDS, i.e.,  $s_i \in R_i^k(\cdot)$ . A subject is then identified as a  $k$ th-order (revealed-)rational player when they exhibit  $m$ th-order revealed rationality for  $m = k$  but not for  $m = k + 1$ . That is, a player is classified into the upper bound of their (revealed) rationality level.<sup>16</sup>

The idea behind the revealed rationality approach is the “as-if” argument: a subject  $i$  selecting  $s_i \in R_i^k(\cdot) \setminus R_i^{k+1}(\cdot)$  in finite observations behaves like a  $k$ th-order rational player, who always selects a strategy in  $R_i^k(\cdot)$  but probably not in  $R_i^{k+1}(\cdot)$ , and thus is identified as a  $k$ th-order revealed rational player. Under this identification criterion, we can identify an individual's order of (revealed) rationality without requiring them to play in multiple games with different payoff structures. In our data analysis, we will classify subjects into five different types: first-order revealed rational (R1), second-order revealed rational (R2), third-order revealed rational (R3), fourth-order (or fully) revealed rational (R4), and non-rational (R0).<sup>17</sup> Tables 1 and 2 summarize the predicted actions under the revealed rationality approach for each type of players in our ring games and guessing games, respectively.

## 5. Experimental design and hypotheses

### 5.1. Treatments

We design a laboratory experiment to measure subjects' higher-order rationality. In the main part of the experiment, subjects first play the ring games, followed by the guessing games, in two different scenarios: the *Robot Treatment* and the *History Treatment*. Using a within-subject design, we alternate the order of the two scenarios (RH Order and HR Order) across sessions to balance out potential spillover effects from one treatment to another.

<sup>16</sup>Kneeland (2015) uses the *exclusion restriction* (ER) as its identification strategy, assuming that a player with low order rationality does not respond to changes in payoff matrices positioned away from herself. However, Lim and Xiong 2016 show that more than three-quarters of their experimental subjects change their actions in two identical ring games, which suggests the failure of the ER assumption since a rational player is predicted to take the same action in two identical games under the exclusion restriction. Also, the ER assumption does not facilitate the identification of higher-order rationality in the guessing games since we cannot separate out first-order payoffs from higher-order ones.

<sup>17</sup>In a four-player ring game, the highest identifiable (revealed) order of rationality is level four.

Table 2. Predicted actions in the guessing games under the revealed rationality approach

Level	Guessing Games		
	$p = 1/3$	$p = 1/2$	$p = 2/3$
R0	34–100	51–100	68–100
R1	12–33	26–50	46–67
R2	5–11	14–25	31–45
R3	2–4	8–13	21–30
R4 (or above)	1	1–7	1–20

In each scenario, each subject first plays the two four-player, three-action ring games (G1 and G2 in Figure 1) in each position in each game once (for a total of eight rounds). Each subject is, in addition, assigned a neutral label (Member A, B, C, or D) before the ring games start. The label is only used for the explanation of an opponent’s strategy in the History Treatment and does not reflect player position. To facilitate the cross-subject comparison, all the subjects play the games in the following fixed order: P1 in G1, P2 in G1, P3 in G1, P4 in G1, P1 in G2, P2 in G2, P3 in G2, and P4 in G2.<sup>18</sup> The order of payoff matrices is also fixed, with a subject’s own payoff matrix being fixed at the leftmost side.<sup>19</sup>

In the Robot Treatment, the subjects play against fully rational computer players. Specifically, each subject in each round is matched with three robot players who only select the strategies that survive iterated dominance elimination (i.e., the equilibrium strategy). We inform the subjects of the presence of robot players that exhibit third-order rationality.<sup>20</sup> The instructions for the robot strategy are as follows:<sup>21</sup>

*When you start each new round, you will be grouped with three other participants who are in different roles. The three other participants will be computers that are programmed to take the following strategy:*

- (1) *The computers aim to earn as much payoff as possible for themselves.*
- (2) *A computer believes that every participant will try to earn as much payoff as one can.*
- (3) *A computer believes that every participant believes “the computers aim to earn as much payoff as possible for themselves.”*

The first line of a robot’s decision rule (“The computers aim to...”) implies that a robot never plays strictly dominated strategies and thus exhibits first-order rationality. The second line (along with the first line) indicates that a robot holds the belief that other players are (first-order) rational and best responds to such belief, which implies a robot’s second-order rationality. The third line (along with the first and second lines) implies that, applying the same logic, a robot exhibits third-order rationality.

<sup>18</sup>Note that Player 4 has a dominant strategy in the ring game. We have our subjects play in each position in the reverse order of the IEDS procedure to mitigate potential framing effects resulting from the hierarchical structure.

<sup>19</sup>This feature is adopted in Jin (2021) and the main treatment of Kneeland (2015). Kneeland (2015) perturbs the order of payoff matrices in a robust treatment and finds no significant effects on subject behavior.

<sup>20</sup>Since level four is the highest identifiable (revealed) order of rationality in a four-player ring game, incorporating a third-order rational computer player is sufficient to identify this maximum level.

<sup>21</sup>Our instructions are adapted from the experiment instructions of Study 2 of Johnson et al. (2002). The original instructions are as follows: “In generating your offers, or deciding whether to accept or reject offers, assume the following: 1. You will be playing against a computer which is programmed to make as much money as possible for itself in each session. The computer does not care how much money you make. 2. The computer program expects you to try to make as much money as you can, and the program realizes that you have been told, in instruction (1) above, that it is trying to earn as much money as possible for itself” (p. 44-45).

In the History Treatment, the subjects play against the data drawn from their decisions in the previous scenario. Specifically, in each round, a subject is matched with three programmed players who adopt actions chosen in the Robot Treatment by three other subjects.<sup>22</sup> Every subject is informed that other human participants' payoffs would not be affected by their choices at this stage. By having the subjects play against past decision data, we can exclude the potential confounding effect of other-regarding preferences on individual actions.

After the ring games, the subjects play the two-person guessing games (in the order of  $p = 2/3, 1/3, 1/2$ ) in both the Robot Treatment and the History Treatment. Instead of being matched with three opponents, a subject is matched with only one player in the guessing games. The instructions for the guessing games in both treatments are revised accordingly.

## 5.2. Hypotheses

The Robot Treatment is designed to convince subjects that the computer opponents they face are the most sophisticated players they could encounter. Consequently, if our Robot Treatment is effectively implemented, it should prompt subjects to employ a strategy at the highest achievable level  $k$ , i.e.,  $k_i(\gamma, \tau_i = \text{Robot}) = c_i(\gamma)$  for all  $\gamma$  and  $i$ . (Recall that  $k_i$  and  $c_i$  denote subject  $i$ 's realized level and capacity, respectively.) This observation gives rise to the first hypothesis we aim to evaluate.

**Hypothesis 1 (Bounded Capacity)**  $k_i(\gamma, \tau_i = \text{History}) \leq k_i(\gamma, \tau_i = \text{Robot})$  for all  $\gamma$ .

In other words, we test whether subjects' rationality levels against robots capture individual strategic reasoning capacity. The corresponding analysis is presented in [Section 6.2](#).

If [Hypothesis 1](#) holds, then we can evaluate several possible restrictions on  $c_i$  by forming hypotheses on  $k_i(\gamma, \text{Robot})$ . In evaluating [Hypothesis 2](#), we examine whether there are stable patterns in (revealed) individual reasoning depth across games.

**Hypothesis 2 (Constant Capacity)**  $k_i(\gamma, \text{Robot}) = k_i(\gamma', \text{Robot})$  for all  $\gamma, \gamma'$ .

This hypothesis imposes the strictest requirement on stability by testing if a player's rationality level remains constant across games. In other words, it assesses whether playing against robots provides a measure of one's *absolute* depth of reasoning. The corresponding analysis is presented in [Section 6.3](#).

In addition to these two hypotheses, Online Appendix B explores two less stringent stability requirements, such as the stability of relative rankings between players' rationality levels (Hypothesis 3) or the consistency of game difficulty in terms of revealed rationality across players (Hypothesis 4).<sup>23</sup>

## 5.3. Cognitive tests

Apart from the ring games and the guessing games, subjects also complete three cognitive tests to measure different aspects of their cognitive ability and strategic reasoning:

- (1) the Cognitive Reflection Test (CRT),
- (2) the Wechsler Digit Span Test, and
- (3) the farsightedness task.

<sup>22</sup>In the HR Order sessions, the choices made by a subject's opponents were drawn from the participants in the Robot Treatment of previous sessions.

<sup>23</sup>Our Hypothesis 2, 3, and 4 correspond to Restriction 2, 3, and 5 in GHW, respectively (see p. 377).

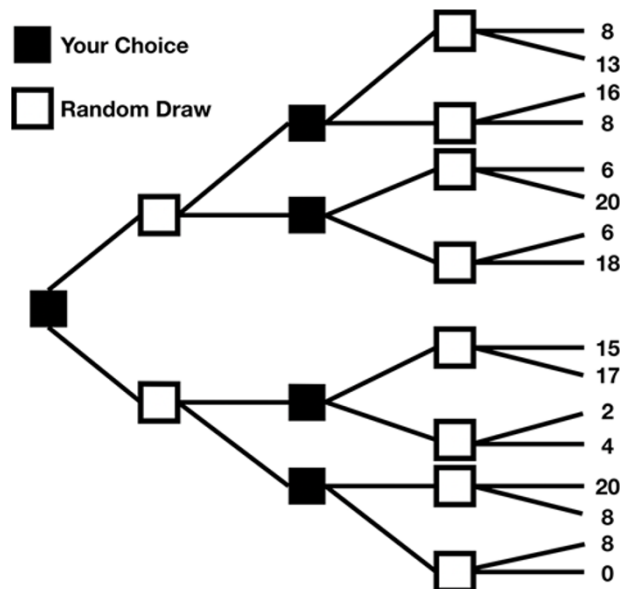


Figure 2. The farsightedness task in Bone et al. (2009)

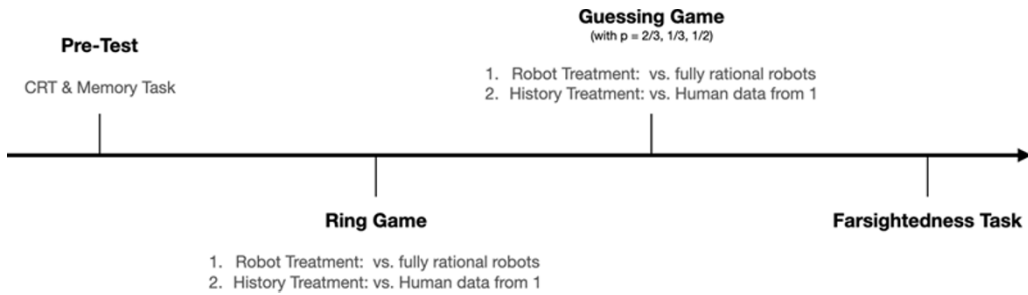
The CRT, proposed by Frederick (2005), is designed to evaluate the ability to reflect on intuitive answers. This test contains three questions that often trigger intuitive but incorrect answers. Performance on this test has been found to be correlated with strategic abilities. For instance, GHW report that the subjects' CRT scores have moderate predictive power on their expected earnings and level- $k$  types.

The second test is the Wechsler Digit Span Test (Wechsler, 1939), which is designed to test short-term memory. In our experiment, this test contains eleven rounds. In each round, a subject needs to repeat a sequence of digits displayed on the screen at the rate of one digit every second. The maximum length of the digit sequence a subject can memorize reflects the subject's short-term memory capacity.<sup>24</sup> Devetag and Warglien (2003) find a positive correlation between individual short-term memory and strategic ability.

Lastly, the *farsightedness task*, developed by Bone et al. (2009), is an individual task to measure a subject's ability to do backward induction, or to anticipate their own future action and make the best choice accordingly. Specifically, it is a sequential task that involves two sets of decision nodes and two sets of chance nodes (see the decision tree in Figure 2). The first and third sets of nodes are the decision nodes where a decision maker is going to take an action (up or down). The second and fourth sets of nodes are the chance nodes where the decision maker is going to be randomly assigned an action (with equal probability).

Notice that there is one dominant action, in the sense of first-order stochastic dominance, at each of the third set of nodes (i.e., the second set of decision nodes). Anticipating the dominant actions at the second set of decision nodes, the decision maker also has a dominant action (down) at the first node. However, if a payoff maximizer lacks farsightedness and anticipates that each payoff will be reached with equal chance, then the dominated action (up) at the first node will become the dominant option from this decision maker's perspective. Therefore, a farsighted payoff-maximizer is expected to choose down, but a myopic one is expected to choose up, at the first move (and choose the dominant

<sup>24</sup>The length of the digit sequence increases from three digits to thirteen digits round by round.



**Figure 3.** Experiment protocol

actions at the second moves). Consequently, we can use their choice at the first move to evaluate the correlation between one's farsightedness and rationality level.

#### 5.4. Laboratory implementation

We conducted 41 sessions between August 31, 2020 and January 28, 2021 at the Taiwan Social Sciences Experiment Laboratory (TASSEL) in National Taiwan University (NTU). The experiment was programmed with the software zTree (Fischbacher, 2007) and instructed in Chinese. A total of 299 NTU students participated in the experiment, all recruited through ORSEE (Greiner, 2015). In our experiment, 136 subjects played the Robot Treatment before the History Treatment in both families of games (RH Order), while 157 subjects played the History Treatment first (HR Order).<sup>25</sup>

Each experimental session lasted about 140 minutes, and the protocol is summarized in Figure 3. At the beginning of the experiment, subjects first completed the CRT and the Wechsler Digit Span Test. After these tasks, subjects played the ring games in both the Robot Treatment and the History Treatment, followed by the guessing games in both treatments. In the final section of the experiment, subjects were asked to complete the farsightedness task. The experimental subjects did not receive any feedback about the outcomes of their choices until the end of the experiment.

There was a 180-second time limit on every subject's decisions in the ring games, guessing games, and farsightedness task. A subject who did not confirm their choice within 180 seconds would have earned zero payoff for that round; however, no subjects exceeded the time limit.<sup>26</sup>

The subjects were paid based on the payoffs (in ESC, Experimental Standard Currency) they received throughout the experiment. In addition to the payoff in the farsightedness task, one round in the ring games and one round in the guessing games were randomly chosen for payment. A subject also got three ESC for each correct answer in the CRT, and one ESC for each correct answer in the Digit Span Test. Including a show-up fee of NT\$200 (New Taiwan dollars; approximately \$7 in USD in 2020), the earnings in the experiment ranged between NT\$303 and NT\$554, with an average of NT\$430.<sup>27</sup>

## 6. Experimental results

In this section, we first provide a general description of the data in Section 6.1. Next, we classify subjects into different rationality levels using the revealed rationality approach in Section 6.2, showing that subjects display higher levels of rationality when playing against robots. In Section 6.3, we demonstrate that individual rationality levels are relatively more stable when controlling for subjects' beliefs about their opponents' depth of reasoning. Finally, we explore the correlation between

<sup>25</sup>Six subjects are dropped from our analysis due to computer crashes.

<sup>26</sup>Jin (2021) sets a 60-second time limit on decisions in the ring games and finds little effect on type classification.

<sup>27</sup>The exchange rate was 1 ESC for NT\$4, and the foreign exchange rate was around US\$1 = NT\$29.4.

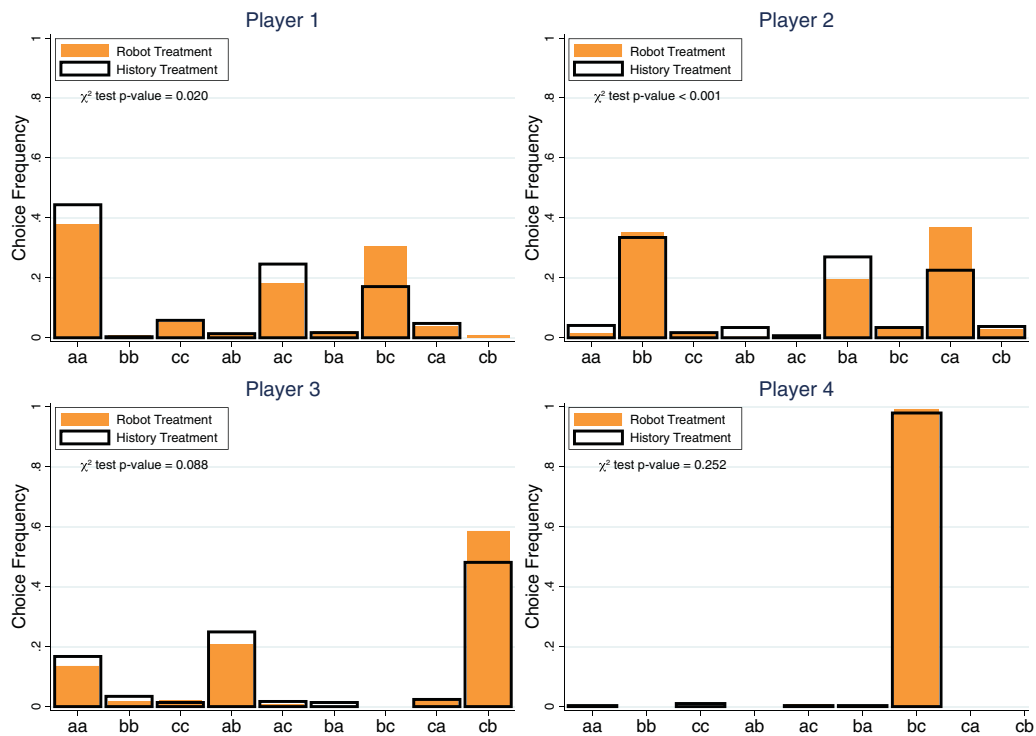


Figure 4. Ring game choice frequency at each position. The first and the second arguments represent the actions of G1 and G2

depth of reasoning, performance on cognitive tests and the heuristics of choosing secure actions in Section 6.4.

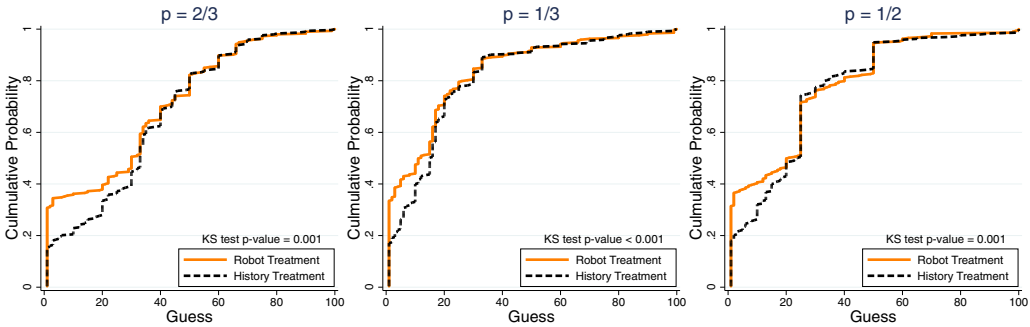
6.1. Data description

Before delving into the main results, we begin by summarizing the subjects' choice frequencies in the ring games (Figure 4) and guessing games (Figure 5). Figure 4 reports the subjects' choice frequencies in the two ring games (G1 and G2, see Figure 1) at each player position. From the figure, we can first observe that in both treatments, over 97% of subjects choose the equilibrium strategy (*b, c*) at P4 ( $\chi^2$  test *p*-value = 0.252). This suggests that subjects are able to recognize strict dominance in the ring games.

Second, at each player position except P4, the significance of  $\chi^2$  tests suggests that subjects' behavior is responsive to the treatments (P1:  $\chi^2$  test *p*-value = 0.020; P2:  $\chi^2$  test *p*-value < 0.001; P3:  $\chi^2$  test *p*-value = 0.088). Moreover, the Robot Treatment shows a 10 to 15 percentage point higher frequency of subjects choosing the equilibrium strategy (*b, c*) at P1, (*c, a*) at P2, and (*c, b*) at P3 compared to the History Treatment, indicating that the Robot Treatment effectively prompts subjects to display higher rationality levels.

Third, at each player position except P4, a notable proportion of subjects choose the secure action that maximizes the minimum possible payoff among the three available actions (*a* at P1, *b* at P2, *a* at P3). As shown in Figure 4, a high proportion of subjects opt for secure actions as an alternative to equilibrium actions. Moreover, except at P4, the proportion of secure actions is higher in earlier positions. At P1, 38% of subjects in the Robot Treatment and 44% in the History Treatment choose the secure actions (*a, a*). This tendency is more pronounced in the History Treatment, where secure actions are





**Figure 5.** Cumulative distribution of guesses

chosen more frequently than equilibrium actions at P1 and P2.<sup>28</sup> This evidence suggests that when players have uncertainty about their opponents' reasoning and strategic behavior, some players may opt for a non-equilibrium strategy to avoid the possibility of experiencing the worst possible payoff.<sup>29</sup> A detailed analysis of the behavior of choosing secure actions is provided in Section 6.4.

Figure 5 presents the cumulative distribution of subjects' guesses across the three guessing games. We observe significant differences in the distributions between the two treatments, regardless of the value of  $p$  ( $p = 2/3$ : KS test  $p$ -value = 0.001;  $p = 1/3$ : KS test  $p$ -value < 0.001;  $p = 1/2$ : KS test  $p$ -value = 0.001). Furthermore, in the Robot Treatment, there is a 13 to 16 percentage point higher proportion of subjects making the equilibrium guess (i.e., choosing 1) across all three guessing games compared to the History Treatment. This difference leads to first order stochastic dominance of the cumulative distribution of guesses in the Robot Treatment over that in the History Treatment, indicating a higher rationality level among subjects in the Robot Treatment for the guessing games.

Furthermore, these distributional differences are driven by variations in equilibrium choices (i.e., 1). After excluding the equilibrium choice of 1, the cumulative distributions between the two treatments are not significantly different for any value of  $p$  ( $p = 2/3$ : KS test  $p$ -value = 0.218;  $p = 1/3$ : KS test  $p$ -value = 0.704;  $p = 1/2$ : KS test  $p$ -value = 0.129). This result further confirms that subjects prompted to perform at their maximum depth of reasoning when facing robots are the primary driving force behind the deeper reasoning observed in the Robot Treatment. In the next section, we will describe our approach for classifying individual rationality levels and perform statistical tests to assess whether subjects demonstrate higher rationality levels when playing against robots.

## 6.2. Rationality level classification

We adopt the revealed rationality approach to classify subjects into different rationality levels. Specifically, let  $s_i = (s_i^\gamma)$  be the vector which collects player  $i$ 's actions in each family of games  $\gamma$ , where  $\gamma \in \{\text{Ring, Guessing}\}$ . In the ring games, we classify subjects based on the classification rule shown in Table 1. In both the Robot Treatment and the History Treatment, if a subject's action profile

<sup>28</sup> In the Robot Treatment at P2, the secure action profile  $(b, b)$  and the equilibrium action profile  $(c, a)$  are chosen 35% and 37% of the time, respectively. By contrast, in the History Treatment, the secure action profile and the equilibrium action profile are chosen 33% and 23% of the time, respectively.

<sup>29</sup> It is also worth noting that at P1 and P2, compared to the Robot Treatment, action profiles involving secure actions in G1 and equilibrium actions in G2 (i.e.,  $(a, c)$  at P1 and  $(b, a)$  at P2) are more frequently observed in the History Treatment. The empirical frequency of action profile  $(a, c)$  at P1 is 18% in the Robot Treatment but 25% in the History Treatment. Similarly, the frequency of action profile  $(b, a)$  at P2 is 19% in the Robot Treatment and 27% in the History Treatment. One potential reason is that choosing  $(a, c)$  at P1 and  $(b, a)$  at P2 are the empirical best response in the History Treatment, and this behavior could be highly rational under a more general notion of rationalizability (Germano et al., 2020). See Online Appendix B for the analysis of the empirical best response in the History Treatment.

matches one of the predicted action profiles of type R0–R4 exactly, then the subject is assigned that level. Therefore, we can obtain each subject's rationality level in the Robot Treatment and the History Treatment, which are denoted as  $k_i(\text{Ring, Robot})$  and  $k_i(\text{Ring, History})$ , respectively.

Similarly, for the guessing games, we classify subjects based on the rule outlined in Table 2. In both treatments, each subject makes three guesses (at  $p = 2/3, 1/3$ , and  $1/2$ ). If a subject is categorized into different levels in different guessing games, we assign the subject the lower level. Thus, we can obtain the levels in both treatments, denoted as  $k_i(\text{Guessing, Robot})$  and  $k_i(\text{Guessing, History})$ , respectively. Following this rationale, we construct the overall distribution of individual rationality levels for each treatment by assigning each subject the lower level they exhibit across the two classes of games, i.e.,  $k_i(\tau_i) = \min\{k_i(\text{Ring}, \tau_i), k_i(\text{Guessing}, \tau_i)\}$ .<sup>30</sup>

Figure 6 reports the overall distribution of rationality levels for the Robot and History Treatments. As shown in the top figure, subjects tend to be classified into higher levels when playing against robots. There are more R1 and R2 players but fewer R3 and R4 players in the History Treatment than in the Robot Treatment. To examine if a subject's reasoning depth is bounded by their revealed rationality level in the Robot Treatment (Hypothesis 1), at the aggregate level, we conduct the two-sample Kolmogorov-Smirnov test to compare the distributions of rationality levels in the two treatments. If Hypothesis 1 holds, we should observe either no difference in the two distributions or the distribution in the Robot Treatment dominating the distribution in the History Treatment. Our results show that the underlying distribution of individual rationality levels in the Robot Treatment stochastically dominates the one in the History Treatment (KS test  $p$ -value = 0.015), and thus provide supporting evidence for Hypothesis 1.

This result is robust across different types of games. As shown in the bottom panels of Figure 6, a similar pattern of first order stochastic dominance is observed regardless of whether rationality levels are classified based on behavior in the ring games or the guessing games (Ring game: KS test  $p$ -value = 0.015; Guessing game: KS test  $p$ -value = 0.001).

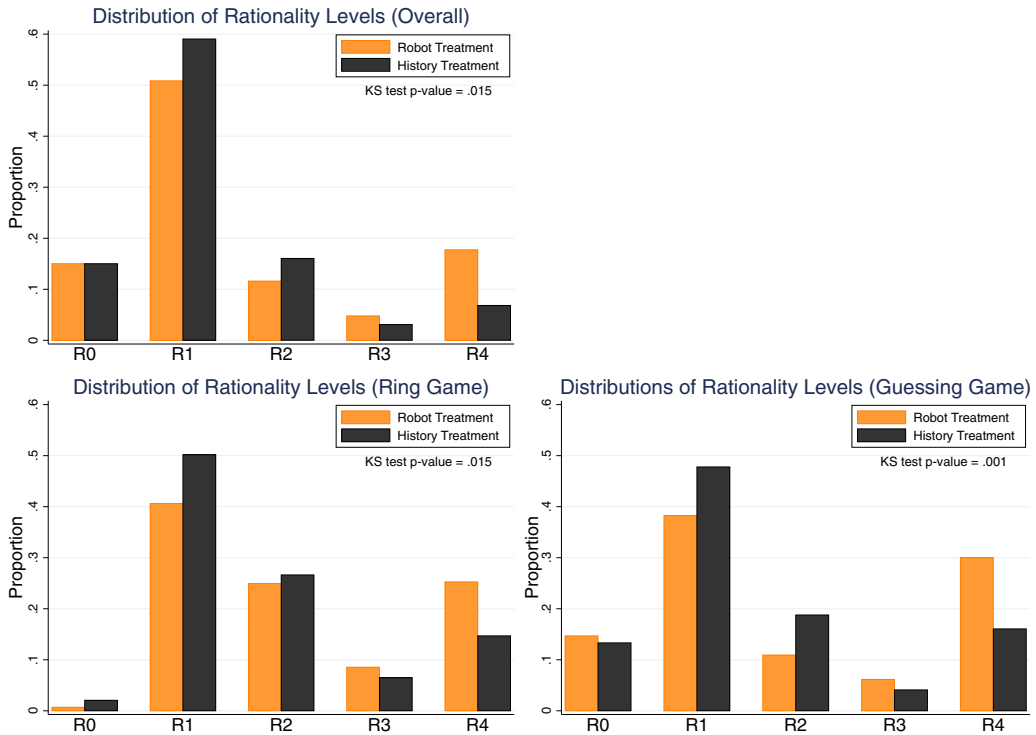
Moreover, our within-subject design gives us paired data of individual rationality levels across treatments, which gives us another way to test Hypothesis 1. Overall, 85 percent of subjects (249/293) exhibit (weakly) higher rationality levels in the Robot Treatment than in the History Treatment. We further conduct the Wilcoxon signed-rank test to examine whether the subjects' rationality levels in the Robot Treatment are significantly greater than the History Treatment. Consistent with Hypothesis 1, we observe higher rationality levels in the Robot Treatment (Wilcoxon test  $p$ -value < 0.0001). Therefore, we conclude that the rationality levels in the Robot Treatment can serve as a proxy of individual strategic reasoning capacity. In Online Appendix B, we separate the data by different games, finding a robust pattern in both.

It is noteworthy that, contrary to previous findings, we observe very few R0 players in the ring games in both treatments (Robot: 0.68%; History: 2.04%).<sup>31</sup> In our experiment, the subjects do not interact with each other in both treatments. Thus, our observation suggests that, when human interactions exist, social preferences may play some roles in a ring game and lead to (seemingly) irrational behavior, though we cannot exclude the possibility that this discrepancy in the prevalence of R0 players is due to different samples.

Yet in the guessing games, our classification results display a typical distribution pattern of estimated levels as documented in Costa-Gomes and Crawford (2006) and GHW. First, the modal type is R1 (Level 1), with more than 35 % of subjects classified as R1 players in both treatments (Robot:

<sup>30</sup> An alternative method for estimating overall levels across games is to impose a probabilistic error structure on deviations from predicted actions (e.g., Stahl & Wilson, 1994; Stahl & Wilson, 1995). However, this is incompatible with the revealed rationality framework, which does not predict a unique best action for each type. Additionally, assigning subjects to the lower order provides a reserved estimate, allowing for a more conservative test when comparing types between the Robot and History Treatments, thus increasing confidence if a statistical difference is observed.

<sup>31</sup> Kneeland (2015) observes 6 % of R0 players (with the ER approach) and Cerigioni et al. 2019 observe more than 15 percent of R0 players (with the revealed rationality approach) in their experiments.



**Figure 6.** Distributions of rationality levels. The top figure is the overall distribution of rationality levels. The bottom figures are the distributions of rationality levels in ring games (Left) and guessing games (Right)

38.23%; History: 47.78%; Costa-Gomes and Crawford (2006): 48.86%; GHW: 50.00%). In particular, the proportion of R1 players reported in the History treatment of our guessing games is very close to the proportion of level-1 players reported in Costa-Gomes and Crawford (2006) and GHW. Second, R3 (Level 3) represents the least frequently observed category among the rational types (i.e., R1–R4), with fewer than 10 percent of subjects classified as R3 players in both treatments, a proportion that aligns with findings in the literature. (Robot: 6.14%; History: 4.10%; Costa-Gomes and Crawford (2006): 3.41%; GHW: 10.34%). Third, the percentage of R4 players in our History Treatment falls within the range of equilibrium-type player proportions reported in Costa-Gomes and Crawford (2006) and GHW (Robot: 30.03%; History: 16.04%; Costa-Gomes and Crawford (2006): 15.91%; GHW: 27.59%). Noticeably, in our Robot Treatment, we observe a relatively high frequency of R4 players compared to previous literature.<sup>32</sup> This finding underscores the significant impact of non-equilibrium belief about opponents on non-equilibrium behavior.

While our subjects' revealed rationality levels are comparatively higher when playing against robots, most do not exhibit more than two steps of reasoning. In the Robot Treatment, around 70 percent of subjects still show an overall rationality level below the third order. This result supports the long-standing idea in the level- $k$  literature: humans have a relatively low cognitive ceiling for strategic thinking, often below level four.

<sup>32</sup>For instance, Arad and Rubinstein (2012) also note that, in their 11–20 money request game, the percentage of subjects employing more than three steps of iterative reasoning does not exceed 20 percent. This aligns with the proportion of R4 players identified in our History Treatment but is lower than that in our Robot Treatment.

Table 3. Markov transition for rationality levels in the robot treatment

From ↓ to →	Guessing Games				
	R0	R1	R2	R3	R4
Ring Games					
R0	<b>50.00% (1)</b>	<b>50.00% (1)</b>	0.00% (0)	0.00% (0)	0.00% (0)
R1	22.69% (27)	<b>45.38% (54)</b>	12.61% (15)	5.88% (7)	13.45% (16)
R2	16.44% (12)	<b>53.42% (39)</b>	6.85% (5)	6.85% (5)	16.44% (12)
R3	8.00% (2)	<b>36.00% (9)</b>	24.00% (6)	0.00% (0)	32.00% (8)
R4	1.35% (1)	12.16% (9)	8.11% (6)	8.11% (6)	<b>70.27% (52)</b>

The number of observations is reported in parentheses.  
The most frequently observed transitions are highlighted in bold.

6.3. Consistency of rationality levels across games

In this section, we evaluate whether controlling for beliefs about the opponent’s depth of reasoning leads individuals to reveal consistent rationality levels across games. There are different notions of consistency. As a first exercise, we assess the strictest form of consistency: an individual reveals *constant* rationality levels across games (Hypothesis 2).

To examine this hypothesis, we generate a Markov transition matrix of rationality levels between the ring games and the guessing games in the Robot Treatment. Table 3 reports the frequency with which an individual moves from each rationality level in the ring games to each rationality level in the guessing games in the Robot Treatment. If the observed individual rationality level is the same across games, then every diagonal entry of each transition matrix in Table 3 will be 100%. Alternatively, if subjects’ rationality levels in the ring games and guessing games are uncorrelated, every row in a transition matrix will be the same and equals the overall distribution in the guessing games.

The transition matrix shows that most R1 and R4 players in the ring games remain as the same level in the guessing games. Most R2 ring game players, however, only exhibit first-order rationality in the guessing games. We do not observe any subjects consistently classified into R3 for both ring and guessing games, possibly because we have relatively low numbers of R3 subjects in either games. Overall, there is a relatively high proportion of subjects (38.23%) that exhibit the same rationality level across games.<sup>33</sup> Also, note that we observe a relatively high proportion ( $52/293 = 17.74\%$ ) of subjects classified as R4 players in both games,<sup>34</sup> suggesting that subjects in our experiment understand the instruction for robots’ decision rules and try to play the best response to such rules.

Moreover, to formally test whether the rationality levels from the ring games and guessing games are independent, we conduct a Monte Carlo simulation in Online Appendix B, following the approach of GHW. Contrary to their findings, we observe that the null hypothesis of independence in rationality levels across games is rejected in the Robot Treatment but not in the History Treatment. Furthermore, we observe that rationality levels in the Robot Treatment are more stable across games than in the History Treatment under alternative notions of consistency compared to Hypothesis 2. These findings provide supportive evidence for our belief control approach.

Although the null hypothesis of independence in rationality levels across games is statistically rejected in the Robot Treatment, the Markov transition matrix reveals that R2 and R3 types appear to be relatively unstable across games. Both R2 and R3 types identified in the ring games tend to “cluster” at R1 in the guessing games, a pattern that is also evident in the History Treatment.

<sup>33</sup> GHW report that only 27.3% of their subjects play at the same level across two families of games.  
<sup>34</sup> In the History Treatment, constant R4 players across games constitute only 6.82% (20/293) of the subjects. See Online Appendix A for the Markov transition matrix for the History Treatment.

**Table 4.** OLS regressions for revealed rationality levels

	Robot Treatment		History Treatment	
	Ring Level	Guess Level	Ring Level	Guess Level
CRT Score	0.298*** (0.072)	0.566*** (0.103)	0.239** (0.074)	0.461*** (0.085)
Memory Score	0.026 (0.036)	0.030 (0.032)	0.005 (0.028)	0.012 (0.034)
Farsightedness	0.569** (0.167)	0.842*** (0.188)	0.339* (0.167)	0.631*** (0.165)
Constant	1.058*** (0.303)	0.092 (0.316)	1.078*** (0.301)	0.187 (0.276)
N	293	293	293	293
R-squared	0.0966	0.1788	0.0556	0.1563

The standard errors are clustered at the session level.

Significance level: \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ .

There are two possible explanations for this pattern. First, the revealed rationality approach identifies an individual's level by the maximum order of rationalizability of their strategy without imposing additional structural assumptions. Since a  $k$ th-order rational player's strategy can also be rationalized by lower-order rationality (see Section 2.2 and Footnote 13), this approach could mechanically “inflate” the identified levels compared to the standard level- $k$  model, which imposes specific structures on beliefs. More importantly, this inflation may be sensitive to the structure of the games, suggesting that the inconsistency of R2 and R3 types could result from this mechanical effect. Second, rather than engaging in higher-order rationality inferences, R2 and R3 types may simply be “avoiding dominated strategies.” To test these hypotheses, in Section 7.3, we compare the revealed rationality approach with the standard level- $k$  model and a diagnostic model, Random, Avoiding Dominated, and Equilibrium (RADE), and report the corresponding results.

#### 6.4. Cognitive tests, secure actions and strategic sophistication

##### 6.4.1. Cognitive tests and strategic sophistication

In addition to analyzing consistency across games, this section investigates whether an individual's performance in other cognitive tests can predict their strategic reasoning ability. To explore this, we regress subjects' revealed rationality levels on their CRT scores, short-term memory task scores, and farsightedness task scores.

The definitions of the independent variables are as follows: *CRT Score* (ranging from 0 to 3) represents the number of correct answers a subject gets in the three CRT questions. *Memory Score* (ranging from 0 to 11) is defined as the number of correct answers a subject provides before making the first mistake. *Farsightedness* is an indicator variable that equals one if a subject chooses to go down at the first move in the farsightedness task (see Section 5.3). Lastly, the dependent variable is the individual revealed rationality levels (ranging from 0 to 4) within each class of games and treatment.

Table 4 presents the OLS regression results for revealed rationality levels. The analysis shows a positive correlation between a subject's CRT performance and their revealed rationality levels across all game types and treatments. Overall, the CRT score is a stronger predictor of rationality levels in the guessing games and the Robot Treatment. In the Robot Treatment, each additional correct answer on the CRT is associated with an average increase of 0.298 ( $p$ -value  $< 0.001$ ) in revealed rationality levels for the ring games, and 0.566 ( $p$ -value  $< 0.001$ ) for the guessing games. In comparison, in the History Treatment, each additional correct answer on the CRT corresponds to a smaller average

increase of 0.239 ( $p$ -value = 0.002) for the ring games and 0.461 ( $p$ -value < 0.001) for the guessing games—approximately 80% of the effect size observed in the Robot Treatment.

In contrast to the previous finding, our results show no significant correlation between short-term memory and strategic sophistication. The coefficient estimates of *Memory Score* are all below 0.03, and all the corresponding  $p$ -values are above 0.3. Notably, these findings are in line with those of GHW, who also observe that CRT scores hold some predictive power over subjects' strategic thinking types, whereas short-term memory capacity does not.

Lastly, an individual's performance on the farsightedness task also significantly predicts their revealed rationality level across all game types and treatments. Similar to the CRT score, we observe a stronger correlation between farsightedness and individual rationality levels in the guessing games and in the Robot Treatment. In the Robot Treatment, a farsighted subject's revealed rationality level is, on average, 0.569 ( $p$ -value = 0.002) and 0.842 ( $p$ -value < 0.001) levels higher than that of a myopic subject when playing ring games and guessing games, respectively. Comparatively, in the History Treatment, a farsighted subject's revealed rationality level is, on average, 0.339 ( $p$ -value = 0.050) and 0.631 ( $p$ -value < 0.001) levels higher than that of a myopic subject when playing ring games and guessing games, respectively. Both of these coefficients are smaller in size compared to the estimates reported for the Robot Treatment. These results indicate a strong correlation between an important strategic thinking skill in a dynamic game—backward induction ability—and the revealed rationality levels in one-shot interactions.

#### 6.4.2. Secure actions in the ring games<sup>35</sup>

Another feature of our modified ring games is that, except at P4, the secure actions differ from the equilibrium actions. This distinction allows us to explore whether players opt for secure actions when they have reached their rationality capacity.

In this section, we analyze the behavior of choosing secure actions by decomposing the revealed rationality levels identified from the ring games into secure and non-secure types. Specifically, for any rationality level  $k$ , a player is classified as  $R_k$ -Secure (or  $R_k$ -S) if they exhibit rationality level  $k$  and choose secure actions in earlier positions.<sup>36</sup> Conversely, a player is classified as  $R_k$ -Non-Secure (or  $R_k$ -NS) if they exhibit rationality level  $k$  but do not choose secure actions in earlier positions. Based on this classification, players are divided into eight possible types: R0, R1-S, R1-NS, R2-S, R2-NS, R3-S, R3-NS, and R4. The distributions for the Robot and History Treatments are shown in Figure 7.

From this figure, we can first observe that there are more secure-type players in the Robot Treatment than in the History Treatment. Among the R1 players, 25.2% are classified as R1-S in the Robot Treatment, while 19.7% are classified as R1-S in the History Treatment. Furthermore, among the R2 players, 42.5% are R2-S in the Robot Treatment, compared to 24.4% in the History Treatment. This result suggests that instead of betting on risky actions, players are more likely to choose a secure action when facing robot players rather than human players.<sup>37</sup>

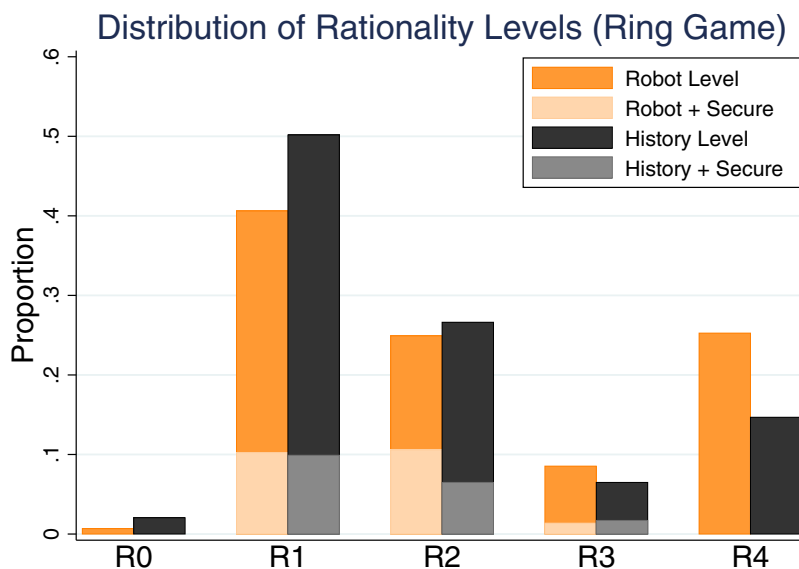
Given this result, we can further explore the behavior of these secure-type players in the guessing games. This is an interesting exercise because there is no secure action in the guessing games, and one might reasonably hypothesize that secure-type players will exhibit higher rationality levels, as their choice of secure actions in the ring games suggests a degree of deliberate, thoughtful decision-making. However, from the Markov transition matrices in Online Appendix A, we find that, in neither the Robot Treatment nor the History Treatment, for any  $k \in \{1, 2, 3\}$ , are the transition probabilities

<sup>35</sup>We thank an anonymous referee for suggesting the analysis of secure actions.

<sup>36</sup>A player is classified as R3-S if they are R3 and choose  $(a, a)$  at P1. Similarly, a player is classified as R2-S if they are R2 and choose  $(a, a)$  at P1 and  $(b, b)$  at P2. A player is classified as R1-S if they are R1 and choose  $(a, a)$ ,  $(b, b)$ , and  $(a, a)$  at P1, P2, and P3, respectively. Note that R1-S corresponds to the type that consistently chooses secure actions.

<sup>37</sup>Refer to Online Appendix B for the joint distribution of rationality levels with secure actions across the Robot Treatment and the History Treatment.





**Figure 7.** Distribution of rational levels with secure actions in the ring games

between  $Rk$ -S and  $Rk$ -NS significantly different.<sup>38</sup> This suggests that the existence of secure actions is indeed a unique feature of our modified ring games. Given any rationality level, choosing secure actions when reaching their rationality capacity does not imply significantly different behavior in the guessing games, where secure actions are absent.

## 7. Discussions

### 7.1. Validity of robot treatment

The validity of the Robot Treatment in eliciting individual strategic thinking capacity relies on our [Hypothesis 1](#) (i.e., that individual rationality levels are higher in the Robot Treatment). An implicit assumption behind this hypothesis is that a subject has an incentive to play at the highest level they can achieve when encountering fully rational opponents playing at their maximum reasoning level. This statement is trivially true for equilibrium-type subjects, as they know their opponents will play the equilibrium strategy and are able to best respond to it. However, for a bounded rational player, this may or may not hold.

If we assume that an iterative reasoning model describes an individual's actual decision-making process, two scenarios explain why a player might only perform  $k$  steps of iterative reasoning. First, they may incorrectly believe that other players can exhibit (at most)  $(k - 1)$ th-order of rationality and best respond to that belief. Second, they may correctly perceive that other players can exhibit (at least)  $k$ th-order of rationality but fail to best respond to it. While our statement regarding incentive compatibility holds in the first case, it becomes unclear how a bounded rational player would respond when facing opponents with rationality levels above  $k$ .

Nevertheless, this scenario does not pose a problem under the identification strategy of the revealed rationality approach. Notice that a player exhibiting  $k$ th-order rationality would also exhibit

<sup>38</sup>To test whether the transition probabilities between  $Rk$ -S and  $Rk$ -NS differ, we conduct  $\chi^2$  tests, with the null hypothesis that the transition probabilities between  $Rk$ -S and  $Rk$ -NS are the same. In the Robot Treatment, the  $p$ -values for R1-S vs. R1-NS, R2-S vs. R2-NS, and R3-S vs. R3-NS are 0.240, 0.338, and 0.582, respectively. Similarly, in the History Treatment, the  $p$ -values for R1-S vs. R1-NS, R2-S vs. R2-NS, and R3-S vs. R3-NS are 0.285, 0.211, and 0.476, respectively.

$m$ th-order rationality for all  $m \leq k$ . Thus, a level- $k$  player  $i$  who perceives other players as exhibiting at least  $k$ th-order rationality also perceives them as exhibiting  $(k - 1)$ th-order rationality. That is, the player knows that their robot opponents' strategies will survive  $k - 1$  rounds of IEDS. Therefore, a payoff-maximizing player  $i$  capable of  $k$  steps of iterative reasoning will choose a strategy in  $R_i^k(\cdot)$ , which contains all undominated strategies after  $k - 1$  rounds of IEDS. Under the revealed rationality approach, player  $i$  will then be classified as a  $k$ th-order revealed-rational player.

Indeed, whether subjects follow the hypothesis and exhibit higher rationality levels when facing fully rational robots is an empirical question. In our setting, we have provided supportive evidence for Hypothesis 1 in Section 6.2. However, it remains an open question whether this increased rationality consistently emerges when individuals encounter robot players in other strategic environments. For instance, in complex games (e.g., Go), individuals might lower their effort and opt for random actions if they perceive highly intelligent robot opponents as unbeatable. Accordingly, exploring how information about robot opponents may influence people's strategic responses in various settings could deepen our understanding of human-robot interactions, especially as AI increasingly shapes human decision-making processes.

## 7.2. Choice of robot strategy instruction

To elicit individual strategic thinking capacity, our Robot Treatment instructions inform subjects that the computer player is third-order rational (i.e., the computer is rational, knows its opponent is rational, and knows its opponent knows it is rational) to control for their beliefs about the sophisticated robot. Previous experimental studies have used different approaches to inform subjects about the strategy of a fully rational, equilibrium robot player, such as explaining the concept of equilibrium (e.g., Costa-Gomes & Crawford, 2006) or fully disclosing the computer player's exact strategy (e.g., Meijering et al., 2012; Hanaki et al., 2016). However, both approaches may introduce a coaching effect: providing background knowledge about the robot's exact strategy or the concept of equilibrium could directly teach subjects how to play and succeed in the specific game, potentially inflating our estimate of their true strategic thinking capacity. On the contrary, we describe the robot players' rationality in a multi-layered, recursive manner without providing specific details about their actions (Johnson et al., 2002), aiming to reduce the risk of over-coaching while still conveying the robot's strategic sophistication.

Despite our efforts, we acknowledge that our instruction strategy might not fully eliminate the possibility of instruction effects, and some subjects might still be influenced by the way the robot players' rationality is described. For instance, some subjects might pick up hints on how to apply the logic of IEDS in our dominance-solvable games, thereby enhancing their depth of strategic thinking. Conversely, others may find the verbal representation of iterative, self-referential logic confusing, which could hinder deeper reasoning. In future experiments, one could evaluate these effects by running a treatment where subjects read the robot instructions but still play against human opponents, then compare their estimated rationality levels to those in a treatment against humans without robot instructions.

Notably, the goal of our design is to capture individual strategic thinking capacity with respect to iterative reasoning by aligning subjects' beliefs about the robot's higher-order rationality. As a result, a limitation of our design is that we do not aim to measure how well subjects form beliefs about the overall distribution of the population's strategic reasoning depth and best respond accordingly, which is a key aspect of strategic sophistication in the sense of Stahl and Wilson (1995)'s "worldly type." An interesting future direction could be to introduce such a "worldly" robot player and examine whether subjects could outsmart this strategically sophisticated type when playing against the robot.<sup>39</sup>

<sup>39</sup>We thank an anonymous referee for encouraging further discussion on our robot strategy instruction.

### 7.3. Diagnosis of the inconsistency of R2 and R3 with alternative models<sup>40</sup>

Following up on the discussion at the end of Section 6.3, we diagnose the inconsistency of R2 and R3 types by comparing the revealed rationality approach with two alternative models: the standard level- $k$  model and a diagnostic model: Random, Avoiding Dominated, and Equilibrium (RADE). The comparison between the revealed rationality approach and the level- $k$  model aims to determine whether the inconsistency arises from a mechanical effect inherent in the revealed rationality approach. Furthermore, the differences between RADE and the revealed rationality approach help clarify whether R2 and R3 types are simply “avoiding dominated strategies” rather than engaging in higher-order rationality inferences.

In our diagnostic analysis, we consider the standard level- $k$  model where level-0 players are assumed to uniformly randomize across all strategies and level- $k$  players best respond to level  $k - 1$  players. The revealed rationality approach, in contrast, does not impose structural assumptions on how irrational players behave or how higher-order rational players respond to lower types. Instead, it classifies individuals based on the highest rationalizable order of their strategy, identifying irrational (R0) players as those who choose strictly dominated strategies. In our ring games, the standard level- $k$  model predicts that a level-1 player best responds to uniformly random opponents by consistently choosing secure actions at all player positions, aligning with the R1-S type (see Footnote 36). Moreover, Player 1’s and Player 2’s secure actions are also best responses to those of Player 2 and Player 3, respectively (see Figure 1). Consequently, level-2 and level-3 players correspond to R2-S and R3-S, respectively, while level 4 (and above) players are classified as R4.<sup>41</sup> In the guessing games, level- $k$  players choose  $50p^k$  for  $k \geq 1$ . To ensure a fair comparison between the revealed rationality approach and the level- $k$  model, we consider five level types: L0, L1, L2, L3, and L4+.

On the other hand, the RADE model can be viewed as a simplified version of Stahl and Wilson (1995), featuring three behavioral types: Random (R), Avoiding Dominated (AD), and Equilibrium (E). Players of the equilibrium type consistently select equilibrium strategies. Consequently, they choose equilibrium actions at all positions in the ring games and select 1 in all three guessing games. Players of the avoiding dominated type never choose a dominated strategy, even though they do not always select equilibrium strategies. Thus, they behave as R1, R2, or R3 in the ring games and avoid strictly dominated numbers in the guessing games. Lastly, random players are assumed to make choices randomly and are the only type that would ever select a strictly dominated strategy.

To evaluate whether the level- $k$  model and the RADE classification offer greater consistency in type classification than the revealed rationality approach, we first classify subjects into revealed rationality types, level- $k$  types and RADE categories based on their behavior in the ring games. These classifications are then used to predict their behavior in the guessing games. Since both the revealed rationality approach and the RADE model generally lack precise predictions for choices in the guessing games, we make quantitative comparisons by assuming that each type of player selects any number consistent with their behavioral type with equal probability.<sup>42</sup> Under this assumption, we evaluate the consistency of the three models by comparing their mean squared deviation (MSD) scores.

<sup>40</sup>We sincerely thank the guest editor, Ido Erev, for suggesting this diagnostic analysis.

<sup>41</sup>The following results remain robust under a more relaxed classification, where a player is classified as level- $k \geq 1$  if their strategy profile deviates from the level- $k$  strategy profile in only one of the eight actions, following the approach of Kneeland (2015) and Jin (2022).

<sup>42</sup>Take the guessing game with  $p = 1/2$  as an example. Under the assumption of uniform randomization, R0 players are assumed to randomize between 51 and 100. R1 players will uniformly randomize between 26 and 50, and so on. In contrast, the RADE model predicts that Random types will uniformly randomize across all numbers, AD types will uniformly randomize between 1 and 50, and E types will choose 1. To ensure a fair comparison of the level- $k$  model’s performance without introducing any parametric noise structure, we assume that a level- $k$  player uniformly randomizes within an interval defined by the midpoints of adjacent levels in the same spirit of Nagel (1995).

Table 5. MSD of revealed rationality, level-*k* and RADE

	Mean Squared Deviation			
	p = 2/3	p = 1/3	p = 1/2	Mean
Robot Treatment				
Revealed Rationality	1101	224	1363	896
Level- <i>k</i>	1060	199	1244	834
RADE	233	170	585	330
History Treatment				
Revealed Rationality	494	212	968	558
Level- <i>k</i>	441	240	931	538
RADE	245	204	706	385

Note: Under the assumption of uniform randomization, all three models provide precise predictions about the percentage of each number chosen. The MSD is computed by summing the squared deviations of these percentages, and the mean is determined by averaging across games.

Table 5 presents the mean squared deviations (MSDs) for the revealed rationality approach, the level-*k* model and the RADE model. When comparing the revealed rationality approach with the level-*k* model, we observe that the MSDs of the level-*k* model are slightly lower than those of the revealed rationality approach. This suggests that the way reasoning levels are classified in the revealed rationality approach only marginally contributes to the inconsistency of the R2 and R3 types.

Furthermore, as shown in the table, the RADE model yields 30%–60% lower MSDs than the other two approaches across all three guessing games in both the Robot and History Treatments. This result suggests that the RADE model achieves greater consistency across games compared to the revealed rationality approach. It also implies that the R2 and R3 types may simply be avoiding dominated strategies rather than engaging in higher-order rationality inferences.<sup>43</sup>

Notably, the MSDs under RADE in the Robot Treatment are smaller than those in the History Treatment, highlighting the effectiveness of our belief control approach. This finding aligns with another notable contrast between the two treatments: the stability of the E types.<sup>44</sup> In the Robot Treatment, 66% of the 74 subjects classified as E types in the ring games remain classified as such in the guessing games. In contrast, in the History Treatment, only 38% of the E types in the ring games retain their classification in the guessing games. Additionally, 58% of the E types in the ring games switch to AD types in the guessing games, suggesting that the inconsistency of intermediate rationality levels may stem from the heuristic of avoiding dominated strategies when faced with a novel game and uncertainty about opponents’ strategies.

In summary, the revealed rationality approach is a classification method that does not rely on any ad hoc assumptions about beliefs and can be universally applied to any dominant-solvable game in the same way. However, these appealing properties come at a cost. As demonstrated in this diagnostic analysis, the consistency of the revealed rationality classification is sensitive to the structure of iterative rationalizable reasoning. Moreover, this consistency can be potentially distorted if players are simply avoiding dominated strategies.

Finally, to further investigate the correlation between the behavior of avoiding dominated strategies and other cognitive abilities, we regress subjects’ RADE types on their performance in three

<sup>43</sup>For other empirical studies challenging the hierarchical reasoning model’s predictive power, refer, for example, to Erev et al. (2015) and Cooper et al. (2024).

<sup>44</sup>Similar to the identification in the ring games, a player is classified as an E type in the guessing games if they always choose 1. A player is identified as an AD type if they are not an E type and never choose a strictly dominated strategy. Lastly, a player is classified as an R type if they have ever chosen a strictly dominated strategy. See Online Appendix A for the joint distributions in both treatments.

**Table 6.** Multinomial logistic regressions for RADE types

	Robot Treatment				History Treatment			
	Ring R	Ring E	Guess R	Guess E	Ring R	Ring E	Guess R	Guess E
	(vs. Ring AD)		(vs. Guess AD)		(vs. Ring AD)		(vs. Guess AD)	
CRT Score	−0.279 (0.323)	0.625* (0.263)	−0.924*** (0.221)	0.900* (0.411)	0.459 (0.392)	0.894* (0.443)	−0.734*** (0.205)	1.319 (0.764)
Memory Score	0.237* (0.111)	0.096 (0.089)	0.112 (0.100)	0.138* (0.065)	−0.075 (0.143)	−0.028 (0.087)	−0.082 (0.086)	0.019 (0.096)
Farsightedness	1.479 (1.265)	0.935** (0.312)	−0.315 (0.590)	1.295*** (0.277)	−0.473 (1.121)	0.781* (0.329)	−0.940 (0.624)	0.660 (0.394)
Constant	−6.509*** (0.929)	−3.809*** (0.966)	−0.146 (0.869)	−4.788*** (1.102)	−4.197* (2.097)	−4.170** (1.472)	0.793 (0.815)	−5.674** (1.939)
N	293		293		293		293	
Pseudo R <sup>2</sup>	0.0741		0.1360		0.0546		0.0940	

The standard errors are clustered at the session level.  
Significance level: \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ .

cognitive tasks. Since RADE type is a categorical dependent variable, we use multinomial logistic regression with the avoid dominated (AD) type as the reference category.<sup>45</sup> Table 6 shows that, whether in the Robot Treatment or the History Treatment, E types identified from the ring games exhibit significantly higher CRT scores and better performance in the farsightedness task than AD types. Additionally, AD types do not show significantly different performance on these cognitive tests compared to R types. In contrast, AD types identified in the guessing games have significantly higher CRT scores than R types. Furthermore, in the Robot Treatment, E types demonstrate significantly better performance across all three cognitive tasks compared to AD types. Yet in the History Treatment, no significant differences in performance are observed between E and AD types. These results complement our main findings in Section 6.4.1 and contrast the two treatments from a different perspective.

## 8. Concluding remarks

This study delves into the cognitive capacity of individuals in strategic interactions. To examine their ability to engage in multi-step reasoning, we conduct an experiment designed to elicit and identify each subject's "rationality bound," while controlling for a subject's belief about their opponent's depth of reasoning. Following the revealed rationality approach, we use two classes of dominance solvable games, ring games and guessing games, as the base games in our experiment. More importantly, to disentangle the confounding impact of beliefs, we introduce equilibrium-type computer players that are programmed to exhibit infinite order of rationality into the experiment. This design allows us to test (1) whether a subject's rationality level is (weakly) higher in the Robot Treatment and (2) whether the observed rationality level in the Robot Treatment exhibits any stable pattern across games.

Overall, our results offer compelling evidence that matching subjects with robot players to elicit and identify individual strategic reasoning ability is an effective approach. First, subjects exhibit a higher rationality level in the Robot Treatment compared to the History Treatment, supporting the hypothesis that a subject plays at their highest achievable rationality level (i.e., their capacity

<sup>45</sup>See Online Appendix A for multinomial logistic regression with the random (R) type as the reference category.

bound) in the Robot Treatment. Second, the observed absolute (and relative) order of rationality in the Robot Treatment appears more stable across different families of games compared to the History Treatment. Adopting the heuristic of avoiding dominated strategies for individual type classification can enhance cross-game stability, yet classification remains more stable in the Robot Treatment than in the History Treatment. Additionally, we find a positive association between a subject's rationality level and their CRT score and backward induction ability, while no significant correlation is observed with short-term memory. These findings indicate that strategic reasoning ability may represent an inherent personal characteristic that is distinct from other cognitive abilities and can be reliably inferred from choice data when subjects' beliefs about others are properly controlled.

Considering that the revealed rationality bound identified in the Robot Treatment can serve as a proxy for an individual's strategic thinking ability, we can independently implement dominance-solvable games, such as ring games and guessing games, with human subjects playing against fully rational computer opponents to effectively elicit and identify human players' strategic capacity, either before or after any lab experiment. By matching human players with computer players, their revealed strategic sophistication is not confounded by their endogenous beliefs about each other's level of sophistication. Furthermore, the robot approach eliminates the need for multiple players to identify a single player's  $k$ th-order rationality in a game, allowing for an individual task that efficiently elicits and identifies a subject's higher-order rationality. Additionally, as the interactions with computer players are independent of the interactions with human players, the two experiences are expected to have minimal influence on each other. Consequently, the measurement of strategic reasoning ability could remain distinct from the behavioral patterns observed in the main experiment session, thereby avoiding any potential contamination between the two.

Ultimately, we believe that such experiment protocol, particularly the robot approach, has the potential to become a standard tool for measuring a player's actual strategic sophistication, analogous to the usage of the established method (for eliciting risk attitude) in Holt and Laury (2002) but applied to the domain of strategic reasoning. By utilizing this tool, we can gain a better understanding of whether non-equilibrium behavior observed in the main experiment can be attributed to bounded strategic thinking capability or other factors, such as non-equilibrium beliefs and social preferences.

As a final remark, note that our robot strategy instruction is designed by progressively revealing layers of the robot's reasoning. By adding or removing these layers, we can introduce a computer player with a higher or lower order of rationality compared to the robot in our experiment, thereby manipulating subjects' beliefs about their opponents' rationality levels. This flexible, layered structure allows the experimental protocol to be more versatile and applicable to a broader range of contexts, rather than being limited to unifying subjects' beliefs. Using this instruction strategy, one could experimentally study, for instance, a player's strategic response and its evolution under different distributions of opponents' rationality levels (Stahl, 1993; Stahl & Wilson, 1995).

**Supplementary material.** The supplementary material for this article can be found at <https://doi.org/10.1017/eec.2025.5>.

**Acknowledgements.** We thank Marina Agranov, Colin F. Camerer, John Duffy, David Gill, Paul J. Healy, Kirby Nielsen, Thomas R. Palfrey, Joseph Tao-yi Wang and audiences at California Institute of Technology, University of Cologne, the 2021 ESA North American Conference and the 2023 Los Angeles Experiments (LAX) Workshop for comments. We are grateful to the editors Roberto Weber and Ido Erev and two referees for their detailed and constructive comments. We thank Wei-Ming Fu for his excellent research assistance. We appreciate Wooyoung Lim's generosity in providing his experimental software for ring games as a reference. We also thank Joseph Tao-yi Wang for his generosity and support of our use of Taiwan Social Sciences Experimental Laboratory (TASSEL) at National Taiwan University. Wei James Chen is supported by Ministry of Science and Technology of Taiwan (MOST 109-2636-H-008-002). Po-Hsuan Lin is supported by National Science Foundation (SES-2243268). This study is approved by National Taiwan University IRB (201910HM015). The experimental design and analysis plan are pre-registered on the Open Science Framework <https://osf.io/gye4u/>. The replication material for the study is available at <https://osf.io/neqxm/> (DOI: 10.17605/OSF.IO/NEQXM). All errors are our own.



## References

- Agranov, M., Potamites, E., Schotter, A., & Tergiman, C. (2012). Beliefs and endogenous cognitive levels an experimental study. *Games and Economic Behavior*, 75(2), 449–463.
- Alaoui, L., Janezic, K. A., & Penta, A. (2020). Reasoning about others' reasoning. *Journal of Economic Theory*, 189, 105091.
- Alaoui, L., & Penta, A. (2016). Endogenous depth of reasoning. *Review of Economic Studies*, 83(4), 1297–1333.
- Alaoui, L., & Penta, A. (2022). Cost-benefit analysis in reasoning. *Journal of Political Economy*, 130(4), 881–925.
- Arad, A., & Rubinstein, A. (2012). The 11–20 money request game a level- $k$  reasoning study. *American Economic Review*, 102(7), 3561–3573.
- Aumann, R. J. (1992). Irrationality in game theory. In: P. Dasgupta, D. Gale, O. Hart, & E. Maskin (Eds.), *Economic Analysis of Markets and Games*, (pp. 214–227). MIT Press.
- Bayer, R. C., & Renou, L. (2016). Logical omniscience at the laboratory. *Journal of Behavioral and Experimental Economics*, 64, 41–49.
- Bernheim, B. D. (1984). Rationalizable strategic behavior. *Econometrica*, 52(4), 1007–1028.
- Bone, J., Hey, J. D., & Suckling, J. (2009). Do people plan?. *Experimental Economics*, 12(1), 12–25.
- Bosch-Rosa, C., & Meissner, T. (2020). The one player guessing game a diagnosis on the relationship between equilibrium play, beliefs, and best responses. *Experimental Economics*, 23(4), 1129–1147.
- Brandenburger, A., Danieli, A., & Friedenberg, A. (2019). Identification of reasoning about rationality [Working paper].
- Burchardi, K. B., & Penczynski, S. P. (2014). Out of your mind eliciting individual reasoning in one shot games. *Games and Economic Behavior*, 84(1), 39–57.
- Cai, H., & Wang, J. T. -Y. (2006). Overcommunication in strategic information transmission games. *Games and Economic Behavior*, 56(1), 7–36.
- Camerer, C. F. (2003). *The 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. *Quarterly Journal of Economics*, 119(3), 861–898.
- Cerigioni, F., Germano, F., Rey-Biel, P., & Zuazo-Garin, P. (2019). Higher orders of rationality and the structure of games [Economics Working Paper No. 1672]. Department of Economics and Business, Universitat Pompeu Fabra.
- Chen, C. -T., Huang, C. -Y., & Wang, J. T. (2018). A window of cognition eyetracking the reasoning process in spatial beauty contest Games. *Games and Economic Behavior*, 111(1), 143–158.
- Chen, W. J., & Krajbich, I. (2017). Computational modeling of epiphany learning. *Proceedings of the National Academy of Sciences*, 114(18), 4637–4642.
- Cooper, D. J., Fatas, E., Morales, A. J., & Qi, S. (2024). Consistent depth of reasoning in level- $k$  models. *American Economic Journal Microeconomics*, 16(4), 40–76.
- Cooper, D. J., & Kagel, J. H.. (2016). Other-regarding preferences a selective survey of experimental results. In H. J. Kagel, & A. E. Roth (Eds.), *The Handbook of Experimental Economics* (Vol. 2, pp. 217–289). Princeton University Press.
- Costa-Gomes, M., & Crawford, V. P. (2006). Cognition and behavior in two-person guessing games an experimental study. *American Economic Review*, 96(5), 1737–1768.
- Costa-Gomes, M., Crawford, V. P., & Broseta, B. (2001). Cognition and behavior in normal-form games an experimental study. *Econometrica*, 69(5), 1193–1235.
- Crawford, V. P., & Iriberri, N. (2007a). Fatal attraction salience, naivete, and sophistication in experimental “hide-and-seek” games. *American Economic Review*, 97(5), 1731–1750.
- Crawford, V. P., & Iriberri, N. (2007b). Level- $k$  auctions can a nonequilibrium model of strategic thinking explain the winner's curse and overbidding in private-value auctions?. *Econometrica*, 75(6), 1721–1770.
- Devetag, G., & Warglien, M. (2003). Games and phone numbers do short-term memory bounds affect strategic behavior?. *Journal of Economic Psychology*, 24(2), 189–202.
- Erev, I., Gilat-Yihyie, S., Marchiori, D., & Sonsino, D. (2015). On loss aversion, level-1 reasoning, and betting. *International Journal of Game Theory*, 44(1), 113–133.
- Fe, E., Gill, D., & Prowse, V. (2022). Cognitive skills, strategic sophistication, and life outcomes. *Journal of Political Economy*, 130(10), 2643–2704.
- Fischbacher, U. (2007). z-Tree Zurich Toolbox for Ready-Made Economic Experiments. *Experimental Economics*, 10(2), 171–178.
- Fong, M. -J., & Wang, J. T. (2023). Extreme (and non-extreme) punishments in sender-receiver games with judicial error an experimental investigation. *Frontiers in Behavioral Economics*, 2, 1096598.
- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4), 25–42.
- Friedenberg, A., Kets, W., & Kneeland, T. (2018). Is bounded rationality driven by limited ability?. [Unpublished paper], W. P. Carey School of Business, Arizona State University.
- Georganas, S., Healy, P. J., & Weber, R. A. (2015). On the persistence of strategic sophistication. *Journal of Economic Theory*, 159, 369–400.
- Germano, F., Weinstein, J., & Zuazo-Garin, P. (2020). Uncertain rationality, depth of reasoning and robustness in games with incomplete information. *Theoretical Economics*, 15(1), 89–122.

- Gill, D., & Prowse, V. (2016). Cognitive ability, character skills, and learning to play equilibrium a level- $k$  analysis. *Journal of Political Economy*, 124(6), 1619–1676.
- Grehl, S., & Tutić, A. (2015). Experimental evidence on iterated reasoning in games. *PLoS One*, 10(8), e0136524.
- Greiner, B. (2015). Subject pool recruitment procedures organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1(1), 114–125.
- Grosskopf, B., & Nagel, R. (2008). The two-person beauty contest. *Games and Economic Behavior*, 62(1), 93–99.
- Hanaki, N., Jacquemet, N., Luchini, S., & Zylbersztejn, A. (2016). Cognitive ability and the effect of strategic uncertainty. *Theory and Decision*, 81(1), 101–121.
- Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, 92(5), 1644–1655.
- Houser, D., & Kurzban, R. (2002). Revisiting kindness and confusion in public goods experiments. *American Economic Review*, 92(4), 1062–1069.
- Jin, Y. (2021). Does level- $k$  behavior imply level- $k$  thinking?. *Experimental Economics*, 24(4), 330–353.
- Jin, Y. (2022). Reinvestigating  $R_k$  behavior in ring games. *Journal of Behavioral and Experimental Economics*, 98, 101878.
- Johnson, E. J., Camerer, C., Sen, S., & Rymon, T. (2002). Detecting failures of backward induction monitoring information search in sequential bargaining. *Journal of Economic Theory*, 104(1), 16–47.
- Kneeland, T. (2015). Identifying higher-order rationality. *Econometrica*, 83(5), 2065–2079.
- Lim, W., & Xiong, S. (2016). On identifying higher-order rationality [Working Paper].
- Lin, P. -H. (2023). Cognitive hierarchies in multi-stage games of incomplete information: theory and experiment. *ArXiv Preprint arXiv:2208.11190v3*.
- Lin, P. -H., & Palfrey, T. R. (2024). Cognitive hierarchies for games in extensive form. *Journal of Economic Theory*, 220, 105871.
- March, C. (2021). Strategic interactions between humans and artificial intelligence lessons from experiments with computer players. *Journal of Economic Psychology*, 87, 102426.
- Meijering, B., Van Rijn, H., Taatgen, N. A., & Verbrugge, R. (2012). What eye movements can tell about theory of mind in a strategic game. *PLoS One*, 7(9), e45961.
- Nagel, R. (1995). Unraveling in guessing games: An experimental study. *American Economic Review*, 85(5), 1313–1326.
- Ohtsubo, Y., & Rapoport, A. (2006). Depth of reasoning in strategic form games. *The Journal of Socio-Economics*, 35(1), 31–47.
- Pearce, D. G. (1984). Rationalizable strategic behavior and the problem of perfection. *Econometrica*, 52(4), 1029–1050.
- Rapoport, A., & Amaldoss, W. (2000). Mixed strategies and iterative elimination of strongly dominated strategies an experimental investigation of states of knowledge. *Journal of Economic Behavior & Organization*, 42(4), 483–521.
- Selten, R. (1991). *Anticipatory Learning in Two-Person Games*. Springer.
- Selten, R. (1998). Features of experimentally observed bounded rationality. *European Economic Review*, 42(3), 413–436.
- Stahl, D. O. (1993). Evolution of Smart <sub>$n$</sub>  players. *Games and Economic Behavior*, 5(4), 604–617.
- Stahl, D. O., & Wilson, P. W. (1994). Experimental evidence on players' models of other players. *Journal of Economic Behavior & Organization*, 25(3), 309–327.
- Stahl, D. O., & Wilson, P. W. (1995). On players' models of other players theory and experimental evidence. *Games and Economic Behavior*, 10(1), 218–254.
- Van den Bos, W., Li, J., Lau, T., Maskin, E., Cohen, J. D., Montague, P. R., & McClure, S. M. (2008). The value of victory social origins of the winner's curse in common value auctions. *Judgment and Decision Making*, 3(7), 483.
- Wang, J. T., Spezio, M., & Camerer, C. F. (2010). Pinocchio's pupil using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games. *American Economic Review*, 100(3), 984–1007.
- Wechsler, D. (1939). *The Measurement of Adult Intelligence*. Williams & Wilkins.