Theory of Mind Ability and Cooperation in the Prisoners Dilemma [*]

Garret Ridinger [†]
University of Nevada, Reno

Michael McBride [‡]
University of California, Irvine

July 2016

Abstract

The ability to accurately assess others' intents, beliefs, and emotions – called Theory of Mind (ToM) – is conjectured to be important for social cooperation. We study the role of ToM ability in fostering cooperation in the simultaneous and sequential prisoners dilemma (PD) games. Our norm-based model predicts that high ToM ability individuals will believe in more cooperation and cooperate at higher rates than low ToM ability individuals in the sequential PD game relative to the simultaneous PD game. Experimental results match these predictions and reveal that ToM ability affects cooperation via beliefs in others' cooperativeness rather than fixed preference traits.

JEL Classification: C73, C92, D03.
Keywords: cooperation, norms, beliefs, theory of mind.

1 Introduction

Humans manifest both highly developed mental abilities and cooperative tendencies (Flinn et al., 2005; Sterelny, 2012; Tomasello, 2014), and researchers have postulated that

[*] We thank workshop participants at UC Irvine for feedback and the Experimental Social Science Laboratory at UC Irvine for use of its facilities. This research was supported by funding from Air Force Office of Scientific Research Award No. FA9550-10-1-0569 and Army Research Office Award No. W911NF-11-1-0332.

[†] Department of Managerial Sciences, College of Business, 1664 N Virginia St., Reno, NV, 89557, gridinger@unr.edu.

[‡] Department of Economics and Experimental Social Science Laboratory, 3151 Social Science Plaza, Irvine, CA, 92697-5100, mcbride@uci.edu, 949-824-7417, 949-824-2182 (fax).

these traits are evolutionarily linked through theory of mind (ToM) (Preston and de Waal, 2002; Flinn et al., 2005). An agent with ToM has the capacity to conceive of others having their own thoughts, intents, beliefs, and emotions (Premack and Woodruff, 1978). ToM typically develops in human children between the ages of 2-4, and, although nearly all individuals develop ToM in their early years, not all individuals manifest the same level of ToM ability upon reaching adulthood (Baron-Cohen et al., 2001). For example, children with autism disorders manifest an inhibited ToM ability (ToM) (Baron-Cohen et al., 2001), and females typically manifest slightly higher ToM ability than men (Kirkland et al., 2013). ToM may be ”the capstone attribute of human cognition” (Robalino and Robson, 2012), but ToM ability is not equally shared across humans.[1]

ToM ability has been postulated to influence social cooperation via two channels. The first channel is beliefs. ToM enables an individual to identify the possible gains from cooperation and then infer others' mental states when making a decision (McCabe et al., 2000). An individual with high ToM ability should more accurately predict the intentions and behavior of another actor and thus be more likely to enjoy the fruits of cooperative endeavors while avoiding exploitation. Higher ToM ability should thus confer a fitness advantage and lead to evolutionary pressures in favor of both high ToM ability and increased social cooperation (Flinn et al., 2005; Sterelny, 2012; Tomasello, 2014). The second channel is through fixed preference traits, though here there is less consensus. The mental capacities that produce ToM are believed to be necessary for empathetic preferences that may lead to increased social cooperation (Preston and de Waal, 2002; Singer and Fehr, 2005; Vollm et al., 2006). If an enhanced ability to understand another's mental and emotional states is correlated with the tendency to feel and care about the other's well-being, then high ToM might be associated with other-regarding preferences and pro-social emotions that increase cooperativeness. However, high ToM ”Machiavellian” agents may be better able to selfishly take advantage of others, thus enhancing their fitness at the expense of other-regarding agents (Whiten and

[1] There is an ongoing debate as to whether any other species has ToM, e.g., see Seyfarth and Cheney (2013).

Byrne, 1997; Maestripieri, 2007).[2] The overall effect of ToM ability on cooperation remains unclear and debated.

We ask: does ToM ability foster social coordination or undermine it, and does it affect cooperation through beliefs or fixed preferences traits? To address this question we conduct a laboratory experiment that places subjects into two different Prisoners Dilemma (PD) settings — one a simultaneous PD game and the other a sequential PD game. In both settings, we record subjects' decisions and elicit their (first-order) beliefs about others' behavior and their (second-order) beliefs about others' beliefs about others' behavior. To obtain a measure of each subject's ToM ability, the subjects complete the Reading the Mind in the Eyes Test (RMET) developed by psychologists and widely used in experimental studies of ToM. We conduct both PD settings because, under a minimal set of assumptions, our norm-based model of utility predicts that the role of ToM ability should differ in the two settings. Specifically, ToM ability should be more positively associated with cooperation in the sequential PD than in the simultaneous PD setting. An added advantage of the sequential setting is that in it we obtain additional measures of beliefs and fixed preference traits, thus allowing us to conduct various tests of the robustness of our results.

We report a number of findings, some of which match our predictions. First, ToM ability is positively correlated with cooperative behavior in the sequential PD game but not in the simultaneous PD game. Second, ToM ability is correlated with more accurate and precise beliefs about others' cooperation in the sequential PD game but not the simultaneous PD setting. Third, ToM ability is positively correlated with higher belief in others' cooperativeness and with beliefs about others' beliefs about cooperativeness. Fourth, ToM ability positively influences cooperation in the sequential PD game through beliefs rather than through fixed preference traits; i.e., there is little evidence that high ToM ability subjects are more inclined to cooperate than low ToM ability subjects once beliefs are controlled.

[2] Research examining the relationship between measures of Machiavellianism and theory of mind has been mixed. Paal and Bereczkei (2007) found no correlation, while Lyons et al. (2010) found that Machiavellianism is negatively correlated with theory of mind. In studies with children, Andreou (2010) and Sutton et al. (2010) have found a positive relationship between ToM skills and Machiavellianism.

This paper provides new perspective for our understanding of cooperation. Our results show that higher ToM ability can lead to increases in cooperation in certain circumstances. While there was no correlation between ToM and decisions to cooperate in the simultaneous prisoner's dilemma, higher ToM ability was associated with higher cooperation and beliefs about the cooperation of others in the sequential prisoner's dilemma. Previous research has shown that conditional cooperation is a commonly chosen strategy in social dilemmas (Fischbacher et al., 2001; Chaudhuri and Paichayontvijit, 2006; Herrmann and Thoni, 2009; Rustagi et al., 2010; Chaudhuri, 2011). However, the success of the conditional cooperation strategy depends on the behavior of others in the population and the ability of conditionalists to recognize cooperators and defectors (Nowak and Sigmund, 2005). In the anonymous, one-shot, simultaneous PD, predicting conditionalists' behavior is difficult because it is unclear what an actor should believe that conditionalist will believe about others' actions. The sequential PD allows for clear conditional strategies as the second mover's action is conditioned on knowledge of the first mover's action, and subjects will higher ToM ability are better able to use their ability to predict what second mover conditionalists will do. It has been argued that conditional strategies may require individuals to possess fairly advanced ToM (Nowak and Sigmund, 2005). We show that individuals who possess higher ToM ability appear better able to recognize how the sequential game will influence the behavior of the population. As a result, higher ToM ability increases cooperation in environments where individuals can use conditional strategies and recognize that others can use conditional strategies.

Economists rarely use the term ”theory of mind;” recent exceptions include Robalino and Robson (2012), Kimbrough et al. (2014), Georganas et al. (2015), and Robalino and Robson (In Press) . However, both macroeconomic and microeconomic theory directly incorporate ToM into its models of human behavior. An example is modern game theory where each actor is depicted as having well-defined preferences and beliefs about other's utility functions and behavior, and an equilibrium is reached when a steady state in beliefs and actions is achieved. Some game-theoretic solution concepts (e.g., rationalizability, iterated elimination of dominated strategies) assume that actors have extremely high ToM ability in the form of precise beliefs about others' behavior, others' beliefs about others' behavior, others' beliefs about others' beliefs about others' behavior, *ad infinitum*. ToM ability also plays a role in the economic theory of cooperation. Cooperation in repeated PD games is achievable when actors have beliefs about others' utilities and others' trigger strategies. Models that postulate social preferences in the form of reciprocity (Bowles and Gintis, 2011), norm following (Bicchieri, 2006; Kessler and Leider, 2012; Kimbrough and Vostroknutov, 2015), inequity aversion (Fehr and Schmidt, 1999; Bolton and Okenfels, 2000), and intentions (Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006) also assume that actors can accurately assess others' intentions, behavior, or well-being. Our study differs from those cited above in that it focuses on the influence of theory of mind on social dilemma games.

There has been relatively little experimental research by economists on the fundamental aspects of ToM directly, but two literatures on ToM in other disciplines are particularly relevant. The first is the psychology literature that uses the RMET to measure ToM ability. A large number of studies have established that the RMET is a reliable measure of ToM ability and have shown it to be correlated strongly with the conceptual notion of ToM (Baron-Cohen et al., 2001; Golan et al., 2006, 2007; Torralva et al., 2007). A few studies by economists have used the RMET (Bruguier et al., 2010; Martino et al., 2013; Ridinger and McBride, 2015; Georganas et al., 2015; Ridinger, 2016), but ours is the first to use the RMET to study the relationship between ToM ability and cooperation. The second literature consists of magnetic imaging studies by neuroscientists that finds neural correlates of ToM. This work shows that the areas the prefrontal regions of the brain are more active in social interactions, including the PD game, thus providing evidence that the PD game is one in which ToM is active (McCabe et al., 2001). These studies do not, however, correlate ToM ability with degrees of cooperativeness.

Of course, there has been a tremendous amount of experimental research by economists on aspects closely related to ToM ability. For example, studies of cognitive hierarchies and level-k behavior examine the strategic component of ToM ability and their results may suggest, similar to the RMET studies, that there exists wide variation in manifested ToM ability (Stahl and Wilson, 1994; Camerer, 2003; Arad and Rubinstein, 2012; Kawagoe and Takizawa, 2012; Georganas et al., 2015). Unlike these studies that focus on strategic reasoning, we use the RMET to measure affective ToM ability, i.e., the ability to understand others' emotional or affective states. The RMET is useful because it provides a direct measure of ToM ability, but we also note that the ability to understand emotions is very relevant for cooperation. Emotions play a role in many social interactions, including social dilemmas like the PD game. Also related to our study are the many experimental studies of behavior in PD games, including those that identify different behavioral types (Fischbacher et al., 2001; Kurzban and Houser, 2005; Herrmann and Thoni, 2009; Fischbacher et al., 2012). As in those studies, we use the strategy method (in our sequential PD treatment) to elicit subjects' behavioral types, but we supplement this strategy method with first and second-order belief elicitation and the RMET measure of ToM ability. Doing so allows us to examine the different channels by which ToM ability influences cooperativeness.

2 Model

2.1 Preferences

A variety of models have been proposed to explain cooperative behavior in social dilemmas such as the PD game. Three prominent ones include the inequity-aversion models, the intentions models, and the norm models. All models distinguish payoff from utility but do so in different ways. Inequity-averion models assume that the individual's utility decreases in payoff inequality (Fehr and Schmidt, 1999; Bolton and Okenfels, 2000). Intentions models assume that an individual's utility depends on what she believes to have been the intent of others (Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006). Norm models assume that an individual's utility decreases when her behavior differs from what she believes she ought to do and what others are expected to do (Bicchieri, 2006; Kessler and Leider, 2012; Kimbrough and Vostroknutov, 2015).

We here use a norm model for two reasons. The first reason is based on experimental evidence from PD games. Although the inequity-aversion and intentions models can account for conditionally-cooperative behavior in PD games, they cannot account for more purely cooperative behaviors and have trouble explaining some behavioral patterns (Clark and Sefton, 2001; Ridinger, 2015). For example, standard versions of these preferences cannot explain, in a sequential PD game, the choice a second mover to cooperate after the first mover is known to defect, a behavior that occurs with non-trivial regularity in experiments (Clark and Sefton, 2001). A norm model can account for this behavior: a second mover with high norm sensitivity may cooperate even after a first mover defection if she believes a norm of cooperation still operates in her community. A norm model thus allows greater flexibility in accounting for various behaviors in PD settings. The second reason is that a norm model has properties relevant to our specific research agenda. Both beliefs about others' behavior and one's own fixed preference trait exist as separate components in the model, and these two channels (beliefs and fixed preferences traits) are the two channels we seek to examine for the influence of ToM on cooperation. A norm model thus allow us to obtain testable predictions about the effect of ToM on cooperation.

Consider the following utility function for an individual:

$$u_i\left(s\right) = \pi_i\left(s\right) - k_i \beta_{hi}^t \left(s_i - \widehat{s}_i\right)^2,$$

where $\pi_i\left(s\right)$ is i's payoff resulting from choice profile s, k_i is i's norm sensitivity (norm salience) that constitutes her fixed preference trait, β_{hi}^t is i's belief about the rate of cooperation among others in the community (i.e., the proportion of others in the community that i believes will adhere to the norm) in treatment t at information set h, s_i is i's decision to cooperate ($s_i = 1$) or defect ($s_i = 0$), and $\widehat{s}_i$ is the norm, i.e., the behavior that i believes she "ought" to do. This specific functional form combines features of functions used in the literature. Like Bicchieri (2006), we assume that the penalty from violating the norm is increasing in the proportion of community members that are believed to adhere to the norm. Like Kessler and Leider (2012), we assume that the penalty is increasing in the difference between the actual action taken and the norm.

We assume that the norm is to cooperate, which gives the following utility function

$$u_i(s) = \pi_i(s) - k_i \beta_{hi}^{t} (s_i - 1)^2 .$$

While this norm assumption makes sense in the sequential prisoner's dilemma, it is possible that the norm may differ in the sequential prisoner's dilemma. For example, the norm for second movers in the sequential prisoner's dilemma could be "cooperate if the first mover cooperates and defect if the first mover defects." Our model can accommodate this assumption, but our main predictions would be largely unchanged. For simplicity, we assume the norm is to always cooperate.

2.2 Simultaneous Prisoners Dilemma

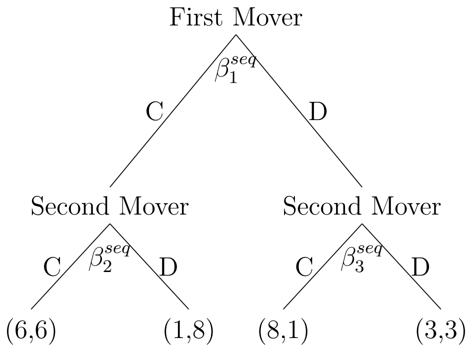
Figure 1(a) shows the simultaneous PD game we will use in our experiment. From a large community of agents, player i and j are selected at random to participate in the simultaneous PD game.

Given the payoffs in the figure and the utility function above, individual i's expected utility from cooperating is

$$\begin{aligned} U_i(s_i = 1, \beta_i) &= \beta_{1i}^{sim}\left(6 - k_i \beta_{1i}^{sim}(1-1)^2\right) + \left(1 - \beta_{1i}^{sim}\right)\left(1 - k_i \beta_{1i}^{sim}(1-1)^2\right) \\ &= 1 + 5\beta_{1i}^{sim}, \end{aligned}$$

		Player 2		
		Cooperate	Defect	
Player 1	Cooperate	6,6	1,8	β_1^{sim}
	Defect	8,1	3,3	$1-\beta_1^{sim}$

(a) Simultaneous



(b) Sequential

Figure 1: Simultaneous and Sequential Prisoner Dilemma Games

and her expected utility from deviating is

$$\begin{aligned} U_i\left(s_i=0,\beta_{1i}^{sim}\right) &= \beta_{1i}^{sim}\left(8-k_i\beta_{1i}^{sim}\left(0-1\right)^2\right)+\left(1-\beta_{1i}^{sim}\right)\left(3-k_i\beta_{1i}^{sim}\left(0-1\right)^2\right) \\ &= 3+5\beta_{1i}^{sim}-k_i\beta_{1i}^{sim}. \end{aligned}$$

Cooperating is the best response for player i when

$$\begin{aligned} U_i\left(s_i=1,\beta_{1i}^{sim}\right) &\geq U_i\left(s_i=0,\beta_{1i}^{sim}\right)\Rightarrow \\ 1+5\beta_{1i}^{sim} &\geq 3+5\beta_{1i}^{sim}-k_i\beta_{1i}^{sim}\Rightarrow \\ k_i\beta_{1i}^{sim} &\geq 2. \end{aligned} \tag{1}$$

Player i is more likely to cooperate when her norm sensitivity is high (large k_i) and when she believes that a large proportion of others will cooperate (large β_{1i}^{sim}). Cooperation depends on both the fixed preference trait k_i and beliefs β_{1i}^{sim}.

2.3 Sequential Prisoners Dilemma

Figure 1(b) shows the sequential PD game. Again there is a large community of agents, and i and j are selected at random to participate in a sequential PD game, with their roles (first or second mover) randomly assigned. Let β_{1i}^{seq} be player i's belief about the proportion of first movers that will cooperate. Let β_{2i}^{seq} and β_{3i}^{seq} be i's belief about the proportion of second movers that will cooperate after the first mover cooperates and defects, respectively.

When making her decision, the second mover knows what the first mover chose, but she also retains a belief about the average behavior of other second movers in the community that affects her norm utility. We assume a sufficiently large community so that i's knowledge of the first mover's action does not affect her belief about the population average level of second mover cooperation.

If the first mover cooperated, then second mover i's utility from cooperating is higher than her utility from defecting when

$$\begin{aligned} U_i\left(s_{2i}=1, s_{1j}=1, \beta_{2i}^{seq}\right) &= 6-k_i\beta_{2i}^{seq}\left(1-1\right)^2 \geq \\ U_i\left(s_{2i}=0, s_{1j}=1, \beta_{2i}^{seq}\right) &= 8-k_i\beta_{2i}^{seq}\left(0-1\right)^2 \Rightarrow \\ k_i\beta_{2i}^{seq} &\geq 2. \end{aligned} \tag{2}$$

Similarly, if the first mover defects, then second mover i cooperates when

$$\begin{aligned} U_i\left(s_i=1, s_{fm}=0, \beta_{3i}^{seq}\right) &= 1-k_i\beta_{3i}^{seq}\left(1-1\right)^2 \geq \\ U_i\left(s_i=0, s_{fm}=0, \beta_{3i}^{seq}\right) &= 3-k_i\beta_{3i}^{seq}\left(0-1\right)^2 \Rightarrow \\ k_i\beta_{3i}^{seq} &\geq 2. \end{aligned} \tag{3}$$

The first mover's expected utility from cooperating is

$$\begin{aligned} U_i\left(s_i=1, \beta_{1i}^{seq}, \beta_{2i}^{seq}, \beta_{3i}^{seq}\right) &= \beta_{2i}^{seq}\left(6-k_i\beta_{1i}^{seq}\left(1-1\right)^2\right)+\left(1-\beta_{2i}^{seq}\right)\left(1-k_i\beta_{1i}^{seq}\left(1-1\right)^2\right) \\ &= 1+5\beta_{2i}^{seq}, \end{aligned}$$

and her expected utility from defecting is

$$\begin{aligned} U_i\left(s_i=0,\beta_{1i}^{seq},\beta_{2i}^{seq},\beta_{3i}^{seq}\right) &= \beta_{3i}^{seq}\left(8-k_i\beta_{1i}^{seq}\left(0-1\right)^2\right)+\left(1-\beta_{3i}^{seq}\right)\left(3-k_i\beta_{1i}^{seq}\left(0-1\right)^2\right) \\ &= 3+5\beta_{3i}^{seq}-k_i\beta_{1i}^{seq}. \end{aligned}$$

Cooperating is player i's best response when

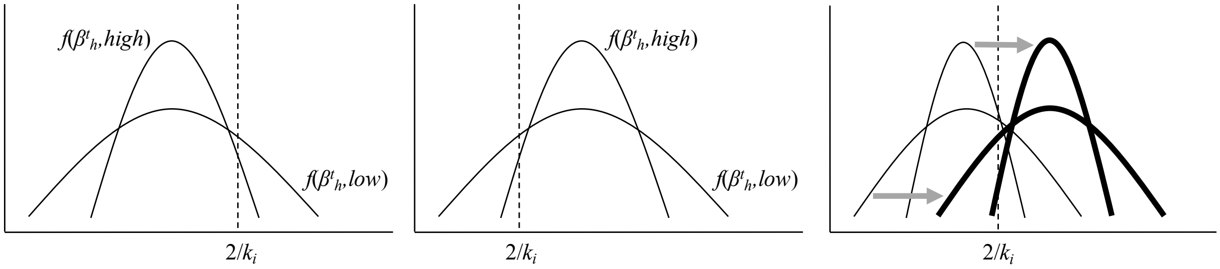
$$\begin{aligned} U_i\left(s_i=1,\beta_{1i}^{seq},\beta_{2i}^{seq},\beta_{3i}^{seq}\right) &\geq U_i\left(s_i=0,\beta_{1i}^{seq},\beta_{2i}^{seq},\beta_{3i}^{seq}\right)\Rightarrow \\ 1+5\beta_{2i}^{seq} &\geq 3+5\beta_{3i}^{seq}-k_i\beta_{1i}^{seq}\Rightarrow \\ k_i\beta_{1i}^{seq}+5\left(\beta_{2i}^{seq}-\beta_{3i}^{seq}\right) &\geq 2\Rightarrow \\ k_i &\geq \frac{2-5\left(\beta_{2i}^{seq}-\beta_{3i}^{seq}\right)}{\beta_{1i}^{seq}}. \end{aligned} \tag{4}$$

This more complicated expression for the first mover's decision is intuitive. When the actor believes others will cooperate (large β_{1i}^{seq}) and has a high norm sensitivity (large k_i), then she is more likely to cooperate. But her beliefs about what others do as second movers also affects her first mover decision. If she believes that second movers are likely to cooperate if she cooperates (large β_{2i}^{seq}), then she is more likely to cooperate as a first mover. If she also believes second movers are more likely to cooperate if she defects (large β_{3i}^{seq}), then she is more likely to defect because the expected payoff from taking advantage of those cooperating second movers is large. Again, both fixed preference traits and beliefs matter.

2.4 Theory of Mind Ability and Cooperation

We now seek predictions for the relationship between ToM ability and cooperation. As mentioned above, there have been diverse conjectures offered in prior literature. We do not seek to privilege one literature over another but instead will derive predictions from a minimal set of assumptions.

Equations (1)-(4) provide the conditions for individual i to cooperate, and in each con-



(a) Beliefs Low K (b) Beliefs High K (c) Beliefs Shift

Figure 2: Examples of Distributional Predictions given Beliefs and ToM

dition ToM ability could potentially affect norm sensitivity k_i or belief β^t_{hi}. Some have speculated that k_i should be positively correlated with ToM ability (Singer and Fehr, 2005), while others have suggested the opposite (Whiten and Byrne, 1997). We are aware of only one empirical paper that measures the propensity to follow norms and ToM Ability, and it finds a positive but only weakly significant relationship between the decision to follow a costly rule and the RMET (Ridinger, 2016). However, that study does not elicit beliefs and therefore cannot distinguish whether ToM ability is correlated with k_i or β^t_{hi} or both. Without additional evidence, we will suppose zero correlation between ToM ability and the norm sensitivity fixed preference trait, and consider this a hypothesis to be tested experimentally.

Hypothesis 1 *Norm sensitivity is independent of ToM ability, $k_i \perp ToM_i$.*

Given that higher ToM ability is defined, in part, as higher belief accuracy, we expect to see beliefs depend on ToM ability. Suppose agents can be partitioned into two types: $ToM_i \in \{high, low\}$. Further suppose that each agent's beliefs β^t_{hi} are drawn from a distribution, with high ToM ability agents drawing from p.d.f. $f\left(\beta^t_h, high\right)$ and low ToM ability agents drawing from $f\left(\beta^t_h, low\right)$. Finally, suppose that these distributions differ in two ways which we call accuracy and precision. High ToM ability agents have higher accuracy so that their mean belief is closer to the true mean cooperation rate than is the mean belief of low ToM ability agents. High ToM ability agents also have greater precision so that the variance in their beliefs is smaller than that of low ToM ability agents. Denote $\widehat{\beta}^t_h\left(ToM\right)$ the mean belief for all i's with ToM ability ToM , $var\left(\widehat{\beta}^t_h\left(high\right)\right)$ the variance in those beliefs across the agents with that same ToM ability, and $\widehat{c}_h^t$ denote the the realized mean cooperation rate. We obtain our second testable hypothesis.

Hypothesis 2 *For each measured belief, high ToM ability subjects have better belief precision and accuracy than low ToM ability agents:*
(i) $\widehat{\beta}_h^t\,(high) - \widehat{c}_i \leq \widehat{\beta}_h^t\,(low) - \widehat{c}_i$*,*
(ii) $var\left(\widehat{\beta}_h^t\,(high)\right) \leq var\left(\widehat{\beta}_h^t\,(low)\right)$*.*

Whether high ToM ability agents are more likely or less likely to contribute than low ToM ability agents will depend on their relative norm sensitivities. Consider the decision in the simultaneous move game and the second mover's decisions in the sequential game. These three decisions have similar conditions for cooperating of the form $\widehat{\beta}_{hi}^t \geq \frac{2}{k_i}$. With low k_i, i.e., high $\frac{2}{k_i}$, as shown in Figure 2(a), it is more likely that a low ToM ability agent will draw beliefs above $\frac{2}{k_i}$ and thus contribute, but with high k_i, as shown in Figure 2(b), it is more likely that the high ToM ability agent will draw beliefs above $\frac{2}{k_i}$. Without knowing the distribution of k_i's we cannot predict whether ToM ability is positively or negatively correlated with cooperation. However, we can conclude that if $f\left(\beta_h^t, high\right)$ and $f\left(\beta_h^t, low\right)$ both shift to the right as depicted in Figure 2(c), then holding k_i fixed we should observe an increase in cooperation for both high and low ToM ability agents but a larger increase in the level of cooperation by those with high ToM ability.

Our experimental treatments are designed with this pattern in mind. Prior studies have found that a sizeable proportion of second movers in sequential PD games play conditional strategies whereby they cooperate if the first mover cooperates but defect if the first mover defects (Bolle and Ockenfels, 1990; Clark and Sefton, 2001; Ahn et al., 2007; Ridinger, 2015). If subject i anticipates that many other subjects will behave reciprocally in this manner, then the distribution of second-mover beliefs after cooperation (β_{2i}^{seq}) should have a higher mean than the distribution of beliefs after defection (β_{3i}^{seq}). The subject should also anticipate higher second-mover cooperation after cooperation when compared to the more difficult to predict simultaneous PD setting.

Hypothesis 3 *Mean beliefs in the sequential PD game will (i) anticipate the presence of reciprocal behavior,* $\widehat{\beta}_2^{seq} \geq \widehat{\beta}_3^{seq}$ *and (ii) anticipate more cooperation by second movers after cooperation than by simultaneous movers,* $\widehat{\beta}_2^{seq} \geq \widehat{\beta}_1^{sim}$*.*

Our next hypothesis follows from the logic on shifting belief distributions described above with the treatment acting as the exogenous factor that shifts the belief distributions.

Hypothesis 4 *The difference in cooperation rates between high and low ToM ability subjects should be (i) higher for second movers after cooperation than for simultaneous movers,*

$$\widehat{c}_2^{seq}\left(high\right) - \widehat{c}_2^{seq}\left(low\right) \geq \widehat{c}_1^{sim}\left(high\right) - \widehat{c}_1^{sim}\left(low\right),$$

which in turn should be (ii) higher than second mover cooperation after defection,

$$\widehat{c}_1^{sim}\left(high\right) - \widehat{c}_1^{sim}\left(low\right) \geq \widehat{c}_3^{seq}\left(high\right) - \widehat{c}_3^{seq}\left(low\right).$$

Now consider the first mover behavior in the sequential PD game. Observe that first movers in the sequential PD game are more likely to contribute than simultaneous movers in the simultaneous game if they believe there are enough reciprocating second movers so that $\widehat{\beta}_2^{seq} - \widehat{\beta}_3^{seq} > 0$. Further notice that because of their enhanced accuracy and precision in beliefs, this is more likely to be true for a high ToM ability subject. It thus follows that condition (4) is more likely to be met for ToM ability subjects than low ToM ability subjects.

Hypothesis 5 *In the sequential PD game, high ToM ability first movers should cooperate at higher rates than low ToM ability first movers, i.e.,*

$$\widehat{c}_1^{seq}\left(high\right) \geq \widehat{c}_1^{seq}\left(low\right).$$

We acknowledge that the predictions above will differ if other assumptions are made about the relationship between norm sensitivity and ToM ability. For example, if the norm sensitivity fixed preference trait is positively correlated with ToM ability, then rates of cooperation should be positively correlated with ToM ability even after controlling for beliefs, while if norm sensitivity is negatively correlated with ToM ability, then the predictions would be quite different with high ToM ability agents possibly cooperating at much lower rates. A richer model of decision making could also be assumed that considers emotions directly. For example, we might predict that high ToM ability individuals would have better accuracy and precision in the sequential move game than the simultaneous move game because they are better able to predict the emotional responses of second movers that might factor into reciprocal behavior. This in turn might lead high ToM ability agents to cooperate at higher rates even if the norm sensitivities do not differ from low ToM ability agents' norm sensitivities. Related reasoning may be used to predict first mover beliefs and behavior to differ for high and low ToM ability agents, e.g., high ToM ability agents may better anticipate more cooperation and thus cooperate at higher rates. We do not rule out these or other possibilities but believe our hypotheses give us a useful starting point for empirical analysis.

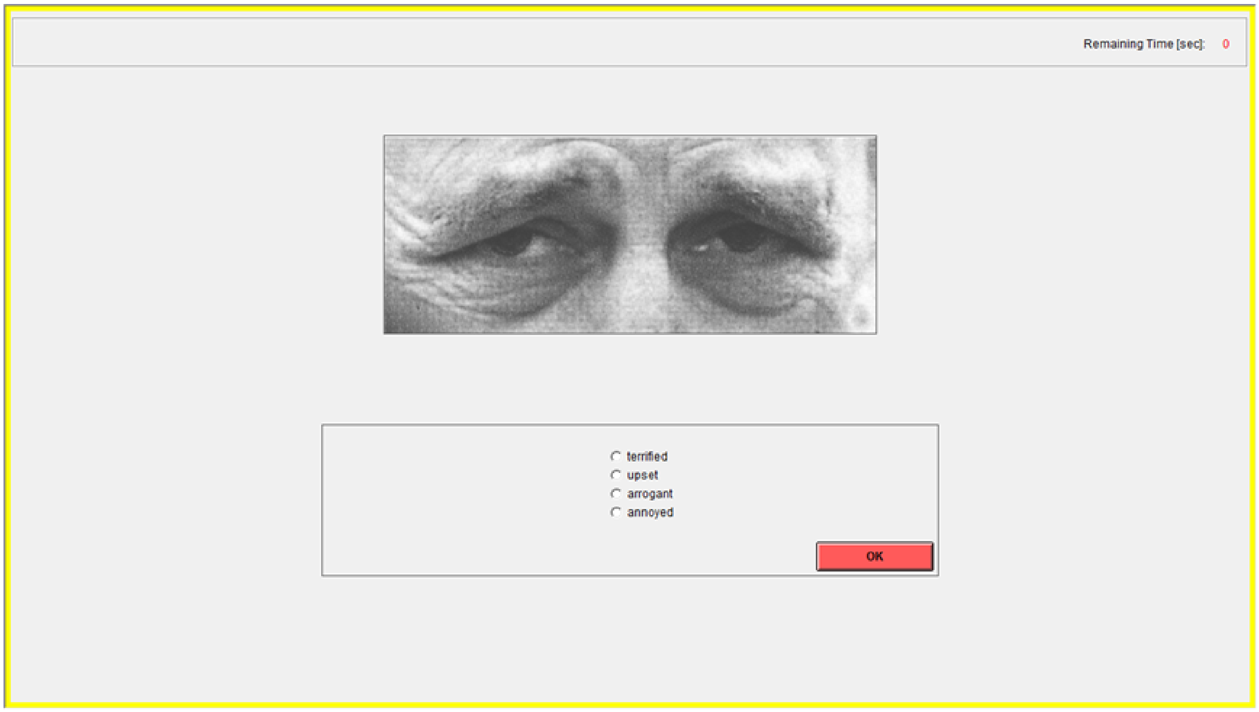
3 Experimental Design

A total of 218 subjects participated in our experiment conducted at the Experimental Social Science Laboratory (ESSL) at the University of California, Irvine. University students learned of the lab via email advertisements and registered to be in the subject pool via an online registration portal. Days before each experiment session, an email is sent to a randomly selected subset of the subject pool notifying them our experiment and providing them a electronic ticket to sign up. Students interested in participating then signed up for a session by clicking the ticket link in the email. Those who signed up received an email reminder the day before the experiment. Subjects were not allowed to participate more than once in the experiment, and there were no other exclusion restrictions for participation other than the student must be at least 18 years of age. All subjects received a show-up payment of $7, plus additional earnings based on their choices and the choices of their randomly-matched partners. The experimental data are available from the authors upon request. This project was approved by the University of California, Irvine Institutional Review Board (HS #2011-8378). To facilitate experimental management, instruction, and data collection, we used the z-Tree software package (Fischbacher, 2007).

Upon arrival at the lab, each subject is randomly placed at one of the lab's computers. After reading a brief instructional screen, each subject participated in a three-stage experimental procedure. Part I is the Reading the Mind in the Eyes Test (Baron-Cohen et al., 2001). Part II is the PD game (simultaneous or sequential). Part III is an extended questionnaire. All decisions and responses are made through the mouse or keyboard. Each session lasted about one hour.

3.1 Reading the Mind in the Eyes Test

The RMET is a 36-question, multiple choice test (Baron-Cohen et al., 2001). For each question, the subject is shown a cropped photograph of the eyes of an individual and, below the photograph, a list of four emotions (see Figure 3). The subject is asked to select which of the four emotions best matches the eyes in the photograph. Each subject is provided a printed handout with a list of definitions of all the emotions and allowed to use this printout when making the selection. The selection is made by using the mouse to select one of the four emotions. After making a selection, the screen goes blank and then advances to the next pair of eyes and four emotions, so each question is asked on its own screen. The set of four emotions for each photograph is different, although some emotions may appear in more than one set. The procedure continues until a selection has been made for each of the 36 pairs of eyes. Subjects were not told how many photographs they would see, although they



Remaining Time [sec]: 0

○ terrified
○ upset
○ arrogant
○ annoyed

Figure 3: Example RMET Question

were told that the entire experiment would not last more than 1.5 hours.[3]

We use the RMET to measure ToM ability instead of other ToM measures for multiple reasons. First, the RMET has been used in a wide number of studies and has shown itself to be a reliable measure of ToM ability (Baron-Cohen et al., 2001; Golan et al., 2006, 2007; Torralva et al., 2007; Kirkland et al., 2013). Second, the RMET produces a measure with meaningful variation in sample sizes like our own that can be exploited in our data analysis (see below). Our subjects' RMET scores range from 16 to 34, with a mean of 27.5, a median of 28, and a standard deviation of 3.32, all similar to other studies. Third, the RMET measures affective ToM ability, and affective ToM ability is useful for our study. Past studies distinguish the affective ToM ability to accurately interpret others' emotions

[3] We implemented an unincentivized RMET. The RMET is typically done in this way, the exception being Ridinger and McBride (2015), who find that incentivizing the RMET negatively affects females' RMET scores but positively affects males' RMET scores. However, this effect is largely a distributional shift. In all regressions, we include a sex dummy as a control variable. As a robustness check we conducted all regressions with interactions between the RMET and sex. The results from the robustness regressions suggest that there is no interaction effect between RMET and sex on the dependent variables. Therefore, we report the non-interacted regressions in the main paper.

from cognitive or strategic ToM to accurately predict other's actions or beliefs, and the two are found to be correlated (Baron-Cohen et al., 2001; Golan et al., 2006, 2007; Torralva et al., 2007). However, by collecting a measure of ToM ability separate from strictly beliefs, we can determine if there is another component of ToM ability separate from beliefs that may affect cooperative preferences. A measure of affective ToM like the RMET is well-suited to this task because it might be supposed that affective ToM might be more likely to be correlated with preferences that are influenced by emotions than would be cognitive or strategic ToM. Such could be the case, for example, if affective ToM is more strongly correlated with empathetic preferences than cognitive or strategic ToM.

3.2 Prisoners Dilemma Game

After completing the RMET, subjects read instructions and participated in one and only one of the two PD games shown in Figure 1. In the Simultaneous treatment, subjects were randomly matched with another participant and asked to choose A ”cooperate” or B ”defect. Subjects were not informed of the other person's choice prior to making their decision. In the Sequential treatment, we used the strategy method. Subjects were randomly matched with another participant, and indicated their choices as both first and second mover. After indicating their choices made, the role was randomly determined by the computer, and the computer then carried out the indicated choices on behalf of the subject. A subject was paid based on the choices she indicated given the role selected by the computer and the choices indicated by her matched partner.

After indicating their choices for the PD game but before learning the results, subjects reported their beliefs about others' actions and beliefs. In the Simultaneous treatment, subjects' beliefs were assessed by asking subjects two questions: (1) What percentage of the other people in the room selected A? and (2) What is the average answer of the other people in the room to Question (1) above. In the Sequential treatment, subjects were asked these two questions for each of the three choices. To incentivize beliefs, we follow Charness and

Dufwenberg (2006) and pay the subjects if her stated belief is within 5 percentage points of the correct belief. If subjects stated a belief within 5 percentage points of the truth, then they received $1. This elicitation procedure is easier for subjects to understand compared to a proper scoring rule and should reduce the frequency of errors in reported beliefs due to confusion. While potentially less confusing for subjects, our belief elicitation procedure is not incentive compatible for beliefs less than 0.05 or greater 0.95.

3.3 Questionnaire

At the end of the experiment, subjects completed a questionnaire that included demographic questions and the cognitive reflection test (CRT) (Frederick, 2005). The CRT has been shown to be highly correlated with intelligence (Toplak et al., 2011). Previous studies have also shown that the RMET score is correlated with intelligence (Baker et al., 2014). Due to this, the CRT will help us control for individual differences in intelligence when using the RMET as a predictor of individual behavior (Ridinger and McBride, 2015).

4 Descriptive Statistics

Tables 1 and 2 present the means and standard deviations for measures used in the subsequent analysis. The average RMET score was approximately 27 and a histogram of the RMET scores can be found in Figure 4. One potential worry is that the RMET distributions may differ across the Simultaneous and Sequential treatments. To investigate this, we conduct a two-sided t-test and find no statistical difference in mean RMET scores between simultaneous and sequential treatments (two-sided t-test, t=0.-1.00, p=0.32). A non-parametric test similarly fails to reject the hypothesis that the RMET scores come from different distributions. (Wilcoxon rank sum test, z=-0.795, p=0.43). This finding lends support to the conclusion that any differential treatment response is more likely due to the treatment effect as opposed to differences in distributions of the RMET between treatments.

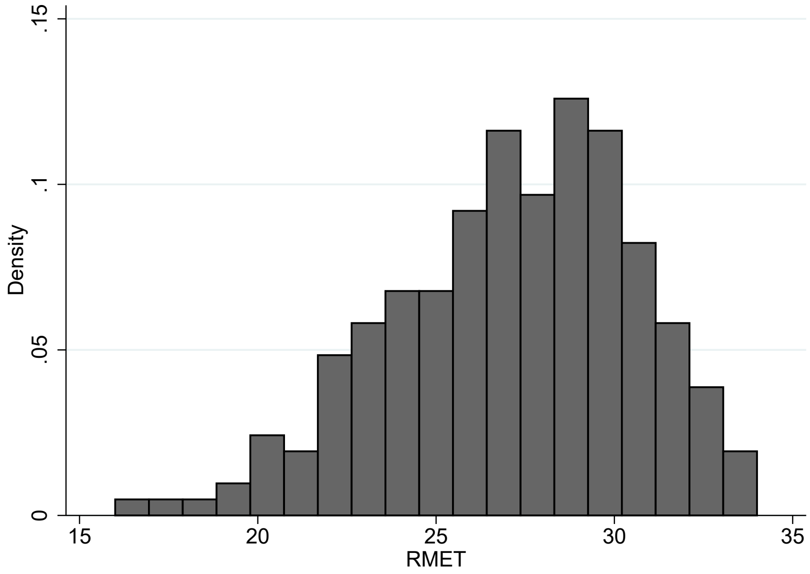


Figure 4: Histogram of RMET Scores

Table 1: Summary Statistics

	Overall mean (sd)	Simultaneous mean (sd)	Sequential mean (sd)
RMET	27.26	27.02	27.51
	(3.62)	(3.88)	(3.32)
Female	0.59	0.68	0.50
	(0.49)	(0.47)	(0.50)
Age	20.09	19.88	20.31
	(1.68)	(1.28)	(2.00)
Native English Speaker	0.50	0.49	0.51
	(0.50)	(0.50)	(0.50)
CRT	1.18	1.19	1.16
	(1.16)	(1.20)	(1.12)
Number of Economic Courses	1.54	1.42	1.66
	(2.63)	(2.40)	(2.86)
Number of Statistics Courses	1.10	1.05	1.14
	(2.63)	(2.40)	(2.86)
Take Home Pay	12.22	12.32	12.12
	(3.32)	(3.91)	(2.55)
Observations	218	112	106

Overall, 59% of the sample was female, and the average age was approximately 20 years old. Subjects earned $12.22 on average.

Table 2: Summary Statistics: Cooperation and Beliefs

	Simultaneous	Sequential
	mean (sd)	mean (sd)
Cooperate	0.49	
	(0.50)	
First Mover Cooperate		0.57
		(0.50)
Second Mover Cooperate		0.62
if First Mover Cooperated		(0.49)
Second Mover Cooperate		0.19
if First Mover Defected		(0.39)
β_1^{sim}, β_1^{seq}	0.50	0.55
	(0.26)	(0.28)
γ_1^{sim}, γ_1^{seq}	0.47	0.52
	(0.22)	(0.25)
β_2^{seq}		0.51
		(0.31)
γ_2^{seq}		0.53
		(0.28)
β_3^{seq}		0.31
		(0.34)
γ_2^{seq}		0.42
		(0.36)
$\lvert\beta_1^{sim} - \bar{s}_1^{sim}\rvert$, $\lvert\beta_1^{seq} - \bar{s}_1^{seq}\rvert$	0.24	0.26
	(0.15)	(0.19)
$\lvert\gamma_1^{sim} - \bar{\beta}_1^{sim}\rvert$, $\lvert\gamma_1^{seq} - \bar{\beta}_1^{seq}\rvert$	0.18	0.22
	(0.14)	(0.18)
$\lvert\beta_2^{seq} - \bar{s}_2^{seq}\rvert$		0.29
		(0.22)
$\lvert\gamma_2^{seq} - \bar{\beta}_2^{seq}\rvert$		0.23
		(0.18)
$\lvert\beta_3^{seq} - \bar{s}_3^{seq}\rvert$		0.30
		(0.23)
$\lvert\gamma_3^{seq} - \bar{\beta}_3^{seq}\rvert$		0.33
		(0.17)
Observations	112	106

Because we elicited second-order beliefs, additional notation is needed. For each subject i, we define the second order belief as γ_h^t where t and h indicate the treatment and information set as before. To measure accuracy of first order beliefs by an individual j, we define the

Table 3: Summary Statistics: Beliefs and ToM

	Simultaneous		Sequential	
	High RMET mean (sd)	Low RMET mean (sd)	High RMET mean (sd)	Low RMET mean (sd)
Cooperate	0.50	0.48		
	(0.50)	(0.50)		
First Mover Cooperate			0.62	0.53
			(0.49)	(0.50)
Second Mover Cooperate			0.70	0.56
if First Mover Cooperated			(0.46)	(0.50)
Second Mover Cooperate			0.19	0.19
if First Mover Defected			(0.40)	(0.39)
β_1^{sim}, β_1^{seq}	0.48	0.51	0.57	0.54
	(0.27)	(0.26)	(0.26)	(0.30)
γ_1^{sim}, γ_1^{seq}	0.48	0.47	0.50	0.53
	(0.22)	(0.23)	(0.23)	(0.27)
β_2^{seq}			0.56	0.47
			(0.28)	(0.33)
γ_2^{seq}			0.56	0.51
			(0.24)	(0.31)
β_3^{seq}			0.38	0.26
			(0.36)	(0.31)
γ_2^{seq}			0.44	0.40
			(0.34)	(0.37)
$\lvert\beta_1^{sim} - \bar{s}_1^{sim}\rvert$, $\lvert\beta_1^{seq} - \bar{s}_1^{seq}\rvert$	0.24	0.24	0.22	0.29
			(0.34)	(0.37)
$\lvert\gamma_1^{sim} - \bar{\beta}_1^{sim}\rvert$, $\lvert\gamma_1^{seq} - \bar{\beta}_1^{seq}\rvert$	0.17	0.19	0.20	0.24
	(0.15)	(0.14)	(0.19)	(0.17)
$\lvert\beta_2^{seq} - \bar{s}_2^{seq}\rvert$			0.24	0.32
			(0.19)	(0.23)
$\lvert\gamma_2^{seq} - \bar{\beta}_2^{seq}\rvert$			0.20	0.26
			(0.18)	(0.18)
$\lvert\beta_3^{seq} - \bar{s}_3^{seq}\rvert$			0.34	0.26
			(0.24)	(0.23)
$\lvert\gamma_3^{seq} - \bar{\beta}_3^{seq}\rvert$			0.34	0.33
			(0.18)	(0.17)
Observations	54	58	47	59

average choice of others as,

$$\bar{s}^t_{hi} = \frac{1}{i-1}\sum_{j\in N/\{i\}} s^t_{hj},$$

where N is the set of subjects in the session. We calculate the accuracy of the first order beliefs for an individual i as $|\beta^t_{hi} - \bar{s}^t_{hi}|$. Similarly, define the average second order belief of others as

$$\overline{\beta}^t_{hi} = \frac{1}{i-1} \sum_{j \in N/\{i\}} \beta^t_{hj},$$

which gives our measure of the accuracy of the second order beliefs as: $|\gamma^t_{hi} - \bar{\beta}^t_{hi}|$.

5 Results

Result 1 *Higher ToM ability is positively associated with cooperation in the sequential PD game but not in the simultaneous PD game.*

Figure 5 shows the average cooperation rates broken down by high and low RMET scores as compared to the median. Cooperation in the Simultaneous treatment is unaffected by differences in the RMET, but in the Sequential treatment, individuals in the high RMET group cooperated at higher rates both as first movers and as second movers after cooperation. Table 4 shows the results of regressions predicting the individual impact of cooperation by RMET score. For the Simultaneous treatment, column (1) shows that the coefficient for RMET in a probit regression is not significant in predicting decisions to cooperate. For the Sequential treatment, column (2) presents an ordered probit regression where the dependent variable is equal to zero if the person defected as both first mover and defected as the second mover when the first mover cooperated ($c^{seq}_1 + c^{seq}_2 = 0$), equal to one if the person either cooperated as a first mover or cooperated as a second mover when the first mover cooperate but not both ($c^{seq}_1 + c^{seq}_2 = 1$), and equal to 2 if the person cooperated as a first mover and cooperated as second mover if the first mover cooperated ($c^{seq}_1 + c^{seq}_2 = 2$). The coefficient on the RMET is significant suggesting that higher RMET increases the probability that an individual will cooperate. This can be seen visually in Figure 6 which plots the predicted probability of each choice in the Sequential treatment by values of the RMET. Higher RMET

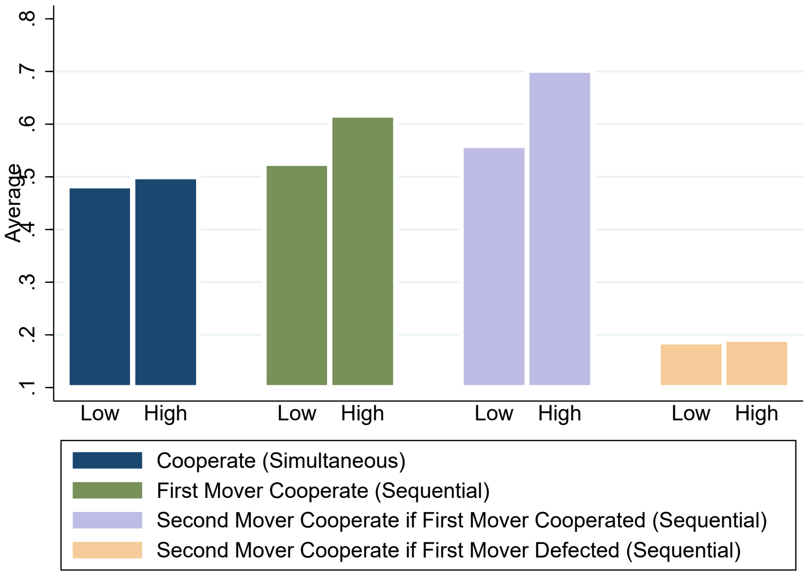


Figure 5: Average Cooperation by Low and High RMET Scores

leads to a greater fraction of individuals who choose to cooperate and a lower fraction who choose to defect at both choices. These results are consistent with Hypotheses 4 and 5, but additional analysis will be done below to identify which channel – beliefs or fixed preference traits (i.e., norm salience) – is the one through which ToM affects cooperativeness.

Result 2 *ToM ability is associated with better accuracy and precision in beliefs in the sequential PD game but not the simultaneous PD game.*

Table 5 presents regression results predicting first-order belief accuracy. In the Simultaneous treatment, there is no significant correlation between RMET and first-order belief accuracy. In the Sequential treatment, individuals with higher RMET scores were significantly more accurate at predicting cooperation by first movers. An increase of one standard deviation in RMET is associated with an approximate 5 percentage point increase in accuracy about first mover cooperation. Similarly, the coefficient for RMET is negative and significant for predicting cooperation of second movers if the first mover cooperates. A one standard deviation increase in RMET is associated with a 4 percentage point increase in accuracy. No significant correlation was found in belief accuracy about cooperation by sec-

Table 4: Predicting Marginal Impact on Cooperation by RMET Score

	(1) Simultaneous Marginal Cooperation Probit	(2) Sequential Marginal Cooperation Ordered Probit
RMET	-0.03	0.09^{**}
	(0.04)	(0.04)
Female	0.87^{***}	0.06
	(0.29)	(0.25)
Age	-0.15	0.10^{*}
	(0.10)	(0.06)
Native English Speaker	-0.23	0.09
	(0.27)	(0.25)
CRT	0.03	-0.23^{**}
	(0.12)	(0.11)
Intercept	3.13	
	(2.10)	
cut 1 cons		3.21^{*}
		(1.65)
cut 2 cons		4.04^{**}
		(1.66)
N	112	104
pseudo R^2	0.10	0.06

Standard errors in parentheses. Regressions in column 1 and 2 are probit and ordered probit, respectively. Regressions include session fixed effects. Dependent variable for column (1) is the dummy variable equal to one if subject chose cooperate. Dependent Dependent variable for column (2) is ordered variable equal to zero if subject defected as first mover and as second mover if first mover cooperated, equal to one if subject cooperated as first mover or as second mover if first mover cooperated, and equal to two if subject cooperated as first and as second mover if first mover cooperated.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

ond movers after first mover defection. As seen in Table 6, a similar pattern is found for second-order belief accuracy.

Hypothesis 2 stated that ToM should be correlated with belief accuracy in both the Simultaneous and Sequential treatments, but our results indicate that affective ToM is only

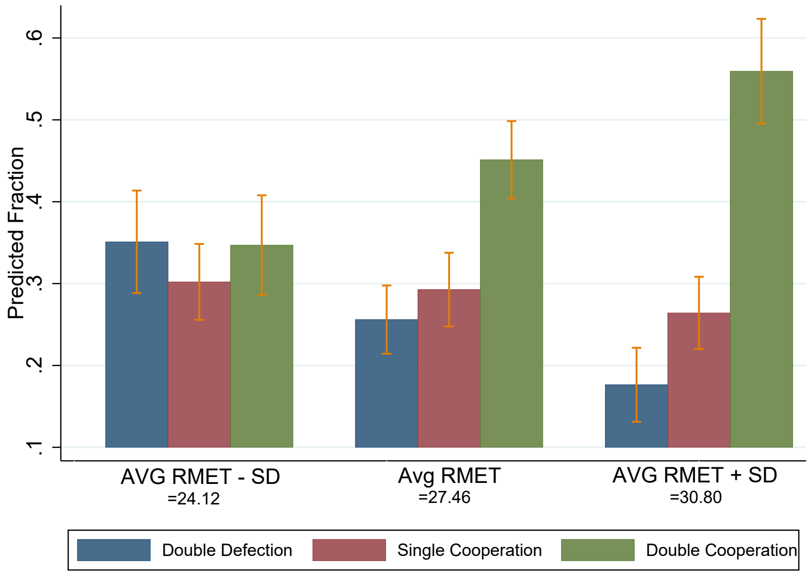


Figure 6: Predicted Distributions of Cooperation Choices in Sequential PD by RMET values

correlated with belief accuracy in the sequential PD Game. One potential explanation for this is that a high-TOM ability individual may be better able to anticipate the presence of reciprocal behavior than low-TOM ability subjects, which leads to more accurate beliefs. The ability of high-ToM ability subjects to leverage their ability appears to be diminished in the simultaneous PD game.

Result 3 *High-ToM ability subjects in the sequential PD game are more likely to believe that cooperation will occur and also that others' believe cooperation will occur.*

Table 7 presents regressions that predicting first-order beliefs by RMET scores. In the Simultaneous treatment, individuals with higher RMET reported lower beliefs about the cooperation rate of others. In the Sequential treatment, higher RMET was significantly associated with holding higher beliefs about first-mover cooperation. An increase of one standard deviation in RMET is associated with a 7 percentage point increase in the belief about first mover cooperation. Similarly, higher RMET is positive and significant in predicting beliefs about second mover cooperation after first-mover cooperated. An increase of one standard deviation in RMET is associate with a 7 percentage point increase in the belief.

Table 5: Predicting First Order Belief Accuracy by RMET Score

	(1)	(2)	(3)	(4)
	Simultaneous		Sequential	
	$\lvert\beta_1^{sim} - \bar{s}_1^{sim}\rvert$	$\lvert\beta_1^{seq} - \bar{s}_1^{seq}\rvert$	$\lvert\beta_2^{seq} - \bar{s}_2^{seq}\rvert$	$\lvert\beta_3^{seq} - \bar{s}_3^{seq}\rvert$
RMET	-0.00	-0.01**	-0.01**	0.00
	(0.00)	(0.01)	(0.01)	(0.01)
Female	-0.00	-0.05	-0.04	-0.06
	(0.04)	(0.04)	(0.05)	(0.05)
Age	0.00	-0.00	0.01	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)
Native English Speaker	0.02	0.02	-0.02	-0.08*
	(0.03)	(0.04)	(0.04)	(0.05)
CRT	0.01	0.01	-0.01	-0.00
	(0.01)	(0.02)	(0.02)	(0.02)
Intercept	0.24	0.65**	0.50*	0.53**
	(0.25)	(0.28)	(0.27)	(0.26)
N	112	104	104	104
R^2	0.032	0.115	0.117	0.100
pseudo R^2				

Robust standard errors in parentheses. OLS regressions include session fixed effects.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

No significant correlation was found between RMET and beliefs about second mover cooperation given the first mover defected. Overall, these results indicate that higher affective ToM leads to better belief accuracy in some settings, in particular, the sequential PD game where ToM ability is better leveraged.

Table 8 shows the regression results on predicting second order beliefs by RMET score. In the Simultaneous treatment, the coefficient on RMET is negative but not significant in predicting beliefs about others' beliefs about cooperation. In the Sequential treatment, RMET is positive but not significant in predicting second order beliefs about first mover cooperation. Higher RMET is significantly associated with higher beliefs about what others'

Table 6: Predicting Second Order Belief Accuracy by RMET Score

	(1)	(2)	(3)	(4)
	Simultaneous	Sequential		
	$\lvert\gamma_1^{sim} - \bar{\beta}_1^{sim}\rvert$	$\lvert\gamma_1^{seq} - \bar{\beta}_1^{seq}\rvert$	$\lvert\gamma_2^{seq} - \bar{\beta}_2^{seq}\rvert$	$\lvert\gamma_3^{seq} - \bar{\beta}_3^{seq}\rvert$
RMET	-0.00	-0.01*	-0.01*	0.01
	(0.00)	(0.01)	(0.01)	(0.01)
Female	-0.02	-0.06	-0.03	-0.03
	(0.03)	(0.04)	(0.04)	(0.04)
Age	0.00	-0.00	0.01	0.00
	(0.01)	(0.01)	(0.01)	(0.01)
Native English Speaker	0.02	0.02	0.00	-0.05
	(0.03)	(0.03)	(0.04)	(0.04)
CRT	-0.00	0.02	0.01	-0.00
	(0.01)	(0.02)	(0.02)	(0.01)
Intercept	0.17	0.53**	0.36	0.15
	(0.26)	(0.24)	(0.25)	(0.29)
N	112	104	104	104
R^2	0.031	0.135	0.078	0.043

Robust standard errors in parentheses. OLS regressions include session fixed effects.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

believe the rate of cooperation will be by second movers when the first mover cooperates. A positive but not significant effect was found for RMET when predicting second order belief about cooperation if the first mover defected.

Hypothesis 3 (i) predicted that on average subjects would anticipate the presence of reciprocal behavior. We can reject the null that the beliefs about cooperation given the first mover cooperated and the beliefs about cooperation given the first mover defected come from the same distribution, suggesting that subjects anticipated reciprocal behavior (Wilcoxon signed-rank test, z=4.99, p=0.00). Hypothesis 3 (ii) suggested that subjects would anticipate more cooperation by second movers after cooperation than by simultaneous movers. We fail to reject that the beliefs about second mover cooperation after cooperation are the same

Table 7: Predicting First Order Beliefs by RMET Score

	(1)	(2)	(3)	(4)
	Simultaneous		Sequential	
	β_1^{sim}	β_1^{seq}	β_2^{seq}	β_3^{seq}
RMET	-0.01*	0.02***	0.02**	0.01
	(0.01)	(0.01)	(0.01)	(0.01)
Female	0.12**	0.07	0.02	-0.04
	(0.06)	(0.06)	(0.07)	(0.07)
Age	0.04*	0.02**	-0.01	-0.02*
	(0.02)	(0.01)	(0.01)	(0.01)
Native English Speaker	-0.01	-0.00	-0.08	-0.15**
	(0.05)	(0.06)	(0.07)	(0.07)
CRT	-0.02	-0.06**	0.00	-0.02
	(0.02)	(0.02)	(0.03)	(0.03)
Intercept	0.15	-0.38	0.17	0.65*
	(0.44)	(0.31)	(0.35)	(0.39)
N	112	104	104	104
R^2	0.118	0.148	0.125	0.077
pseudo R^2				

Robust standard errors in parentheses. OLS Regressions include session fixed effects.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

as the beliefs about cooperation in the Simultaneous treatment (Wilcoxon rank sum test, z=-0.504, p=0.61).

Result 4: *ToM ability positively influences cooperation in the sequential PD game primarily through beliefs rather than through preferences; there is little evidence that high-ToM ability subjects are more inclined to cooperate than low-ToM ability subjects once beliefs are controlled.*

Table 9 presents regression results predicting cooperation by RMET score. We find that RMET score does not predict cooperation in the Simultaneous treatment, but it does in the sequential PD game for first movers and second movers after cooperation. However, as

Table 8: Predicting Second Order Beliefs by RMET Score

	(1)	(2)	(3)	(4)
	Simultaneous	Sequential		
	γ_1^{sim}	γ_1^{seq}	γ_2^{seq}	γ_2^{seq}
RMET	-0.01	0.01	0.03***	0.01
	(0.01)	(0.01)	(0.01)	(0.01)
Female	0.11**	0.01	0.01	-0.03
	(0.05)	(0.05)	(0.06)	(0.08)
Age	0.02	0.02*	0.02*	0.00
	(0.02)	(0.01)	(0.01)	(0.02)
Native English Speaker	-0.00	-0.02	-0.09	-0.17**
	(0.05)	(0.05)	(0.06)	(0.08)
CRT	0.02	-0.06**	-0.03	-0.04
	(0.02)	(0.02)	(0.03)	(0.03)
Intercept	0.19	-0.08	-0.65**	0.01
	(0.38)	(0.30)	(0.32)	(0.51)
N	112	104	104	104
R^2	0.087	0.107	0.185	0.097

Robust standard errors in parentheses. OLS regressions include session fixed effects.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

shown in Table 10, once we control for beliefs the coefficient on RMET is no longer significant. This is consistent with our Hypothesis 1 which suggested that the individual, fixed preference trait propensity to follow norms, after controlling for beliefs, is not influenced by affective ToM. Subjects with higher RMET held higher beliefs about cooperation in the sequential PD game, and that is why they cooperated at higher rates.

Table 9: Predicting Cooperation by RMET

	(1)	(2)	(3)	(4)
	Simultaneous	Sequential		
	Cooperate	First Mover Cooperate	Second Mover Cooperate if First Mover Cooperated	Second Mover Cooperate if First Mover Defected
RMET	-0.01	0.03*	0.03**	0.01
	(0.01)	(0.01)	(0.02)	(0.01)
Female	0.31***	0.11	-0.07	0.02
	(0.10)	(0.11)	(0.10)	(0.07)
Age	-0.05	0.04**	0.01	-0.03*
	(0.04)	(0.02)	(0.01)	(0.02)
Native English Speaker	-0.08	0.01	0.02	-0.19**
	(0.10)	(0.10)	(0.10)	(0.08)
CRT	0.01	-0.05	-0.09**	-0.05*
	(0.04)	(0.05)	(0.04)	(0.03)
Intercept	1.54**	-0.70	-0.36	0.76*
	(0.74)	(0.55)	(0.52)	(0.44)
N	112	104	104	104
R^2	0.136	0.102	0.113	0.226

Robust standard errors in parentheses. OLS regressions include session fixed effects.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 10: Predicting Cooperation by RMET and Beliefs

	(1)	(2)	(3)	(4)
	Simultaneous	Sequential		
	Cooperate	First Mover Cooperate	Second Mover Cooperate if First Mover Cooperated	Second Mover Cooperate if First Mover Defected
RMET	0.00	0.00	0.02	0.01
	(0.01)	(0.01)	(0.01)	(0.01)
β_1^{sim}, β_1^{seq}	0.96***	1.07***		
	(0.14)	(0.13)		
β_2^{seq}-β_3^{seq}		0.00**		
		(0.00)		
β_2^{seq}			0.53***	
			(0.16)	
β_3^{seq}				0.43***
				(0.10)
Female	0.20**	0.03	-0.08	0.04
	(0.10)	(0.08)	(0.09)	(0.07)
Age	-0.08***	0.01	0.02	-0.02
	(0.03)	(0.02)	(0.02)	(0.01)
Native English Speaker	-0.07	-0.00	0.07	-0.13*
	(0.09)	(0.08)	(0.09)	(0.08)
CRT	0.02	0.00	-0.09**	-0.04*
	(0.04)	(0.04)	(0.04)	(0.03)
	1.40**	-0.21	-0.45	0.48
	(0.55)	(0.43)	(0.49)	(0.37)
N	112	104	104	104
R^2	0.358	0.481	0.213	0.356

Standard errors in parentheses. Regressions include session fixed effects.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 11: Predicting RMET by Dictator and Rule-following Task Behavior

	(1)	(2)	(3)
	RMET	RMET	RMET
Donation	0.16		0.12
	(0.25)		(0.26)
Waiting Time		0.05	0.05
		(0.07)	(0.07)
Female	-0.64	-0.75	-0.86
	(1.03)	(0.95)	(1.02)
Age	-0.09	-0.15	-0.13
	(0.27)	(0.28)	(0.27)
Native English Speaker	1.34	1.36	1.39
	(0.90)	(0.89)	(0.90)
CRT	-0.28	-0.21	-0.21
	(0.41)	(0.44)	(0.44)
Intercept	29.04***	29.47***	28.94***
	(5.89)	(6.04)	(5.87)
N	77	77	77
R^2	0.048	0.056	0.059

Robust standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 12: Predicting Correct Choice by RMET Score

	(1)	(2)
	Correct if First Mover Cooperated	Correct if First Mover Defected
RMET	0.52***	0.12
	(0.10)	(0.23)
Female	0.56	-0.80
	(0.87)	(1.67)
Age	-0.41*	0.50
	(0.24)	(0.60)
Native English Speaker	-0.15	-0.60
	(0.89)	(1.76)
CRT	0.05	0.46
	(0.38)	(0.70)
Intercept	18.13***	16.15
	(5.90)	(15.95)
N	77	77
R^2	0.223	0.031

Robust standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

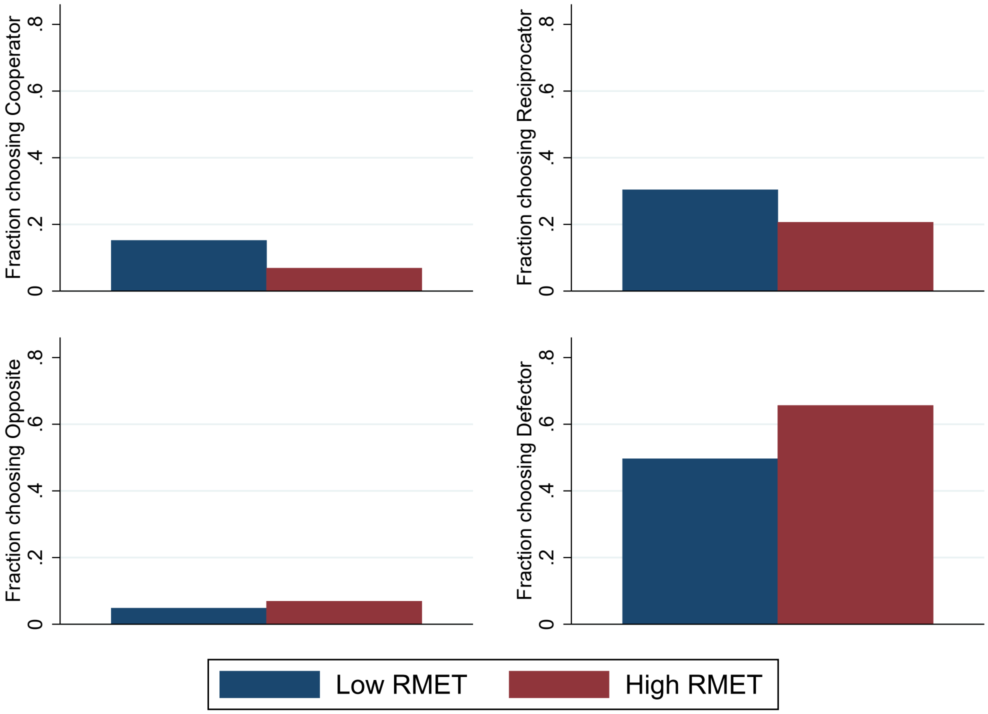


Figure 7: Predicted Choice Types for Negative Emotions by RMET values

6 Conclusion

Our study set out to understand how differences in ToM ability influence cooperation. We introduced a theoretical framework that reveals the effect of ToM ability on cooperation to be generically ambiguous, but through appropriate experimental design we can manipulate subjects beliefs to obtain a differential effect on beliefs according to ToM ability. In particular, we predicted that ToM ability will have a larger impact on cooperation and beliefs about others' cooperation in the sequential PD game relative to the simultaneous PD game. Our laboratory experiment yielded results that confirmed this conjecture and further showed that the effect of ToM ability operates through beliefs rather than fixed preferences traits.

One main implication of our study is that, contrary to the conjectures in existing literature, we find that ToM ability and fixed social preference traits are independent. As explained earlier, researchers have disagreed on the matter. Some suggested a positive link while others have suggested a negative one, but we find no such link. The effect of ToM ability on cooperation is largely through beliefs rather than fixed preference traits, and this in turn implies that whether or not ToM ability helps or hurts cooperation will depend heavily on features of social dilemma setting. In a setting like our sequential PD game where many individuals play conditional strategies, having high ToM ability increases the willingness of first movers to initiate cooperation as their high ToM ability better enables the first movers to foresee second-mover conditional cooperation and their belief that many other first movers will cooperate improves the norm payoff. However, if the distribution of subjects' strategies were different with a much lower level of conditional cooperation, then high-ToM ability subjects would be less likely to cooperate. ToM ability thus improves cooperation only in the right setting, and it happened that our sequential PD game was one such setting.

Our study speaks to but, of course, does not resolve the debate in the literature about the evolution of ToM and cooperation. Relative to early humans or our closest non-human primate relatives, our sample of university student subjects includes only decision makers with fairly high ToM ability. Indeed, the RMET measure could be understood as identifying individual differences in ToM ability among the far right tail of the ToM ability distribution that considers all species. Recognition of this fact perhaps makes our results more striking. That we find differences in cooperativeness by ToM ability even among our set of relatively high-ToM ability actors suggest that even larger differences in ToM ability may have played an important role in the evolution of our species. Our experiment can only offer hints at this larger question, but our findings do suggest that ToM ability might best be understood as a trait that is not directly correlated with all other factors that determine social preferences. Future theoretical work may take advantage of that fact.

There is also much room for additional experimental work. There are, for example, a wide range of strategic scenarios in which ToM may matter greatly, and future work should investigate in which of them ToM does indeed matter. We hope that the lessons of our work will inform this future work.

References

Ahn, T., Lee, M., Ruttan, L., Walker, J., 2007. Asymmetric payoffs in simultaneous and sequential prisoner's dilemma games. Public Choice 132, 353–366.

Andreou, E., 2010. Bully/victimproblems and their association with machiavellianism and self-efficacy in greek primary schools children. Educational Psychology.

Arad, A., Rubinstein, A., 2012. The 11-20 money request game: A level-k reasoning study. American Economic Review 102 (7), 3561–3573.

Baker, C. A., Peterson, E., Pulos, S., Kirkland, R. A., 2014. Eyes and iq: A meta-analysis of the relationship between intelligence and "reading the mind in the eyes". Intelligence 44, 78–92.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., Plumb, I., 2001. The "reading the mind in the eyes" test revised version: A study with normal adults, and adults with asperger syndrome or high-functiong austism. Journal of Child Psychology and Psychiatry 42, 241–251.

Bicchieri, C., 2006. The Grammar of Society: The Nature and Dynamics of Social Norms. Cambridge University Press.

Bolle, F., Ockenfels, P., 1990. Prisoners' dilemma as a game with incomplete information. Journal of Economic Psychology 11, 69–84.

Bolton, G. E., Okenfels, A., March 2000. Erc: A theory of equity, reciprocity, and competition. The American Economic Review 90 (1), 166–193.

Bowles, S., Gintis, H., 2011. A Cooperative Species: Human Reciprocity and Its Evolution. Princeton University Press.

Bruguier, A. J., Quartz, S. R., Bossaerts, P. L., 2010. Exploring the nature of "trading intuition". Journal of Finance 65, 1703–1723.

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

Charness, G., Dufwenberg, M., 2006. Promises and partnership. Econometrica 74, 1579–1601.

Chaudhuri, A., 2011. Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. Experimental Economics 14, 47–83.

Chaudhuri, A., Paichayontvijit, T., 2006. Conditional cooperation and voluntary contributions to a public good. Economics Bulletin 3, 1–14.

Clark, K., Sefton, M., 2001. The sequential prisoner's dilemma: Evidence on reciprocation. The Economic Journal 111, 51–68.

Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games and Economic Behavior 47, 268–298.

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

Fehr, E., Schmidt, K. M., August 1999. A theory of fairness, competition, and cooperation. The Quarterly Journal of Economics 114 (3), 817–868.

Fischbacher, U., 2007. z-tree: Zurich toolbox for ready-made economic experiments. Experimental Economics 10, 171–178.

Fischbacher, U., Gachter, S., Fehr, E., 2001. Are people conditionally cooperative? evidence from a public goods experiment. Economics Letters 71, 397–404.

Fischbacher, U., Gachter, S., Quercia, S., 2012. The behavioral validity of the strategy method in purlic good experments. Journal of Economic Psychology 33 (4), 897–913.

Flinn, M. V., Geary, D. C., Ward, C. V., 2005. Ecological dominance, social competition, and coalitionary arms races: Why hhuman evolved extraordinary intelligence. Evolution and Human Behavior 26, 10–46.

Frederick, S., 2005. Cognitive relection and decision making. Journal of Economic Perspectives 19, 25–42.

Georganas, S., Healy, P. J., Weber, R. A., 2015. On the persistence of strategic sophistication. Journal of Economic Theory 159, 369–400.

Golan, O., Baron-Cohen, S., Hill, J., 2006. The cambridge mindreamind (cam) face-voice battery: Testing complex emotion recognition in adults with and without asperger syndrome. Journal of Autism and Developmental Disorders 36 (2), 169–83.

Golan, O., Baron-Cohen, S., Hill, J., Rutherford, M., 2007. The 'reading the mind in the voice' test-revised: a study of complex emotion recognition in adults with and without autism spectrum conditions. Journal of Autism and Developmental Disorders 37 (6), 1096–1106.

Herrmann, B., Thoni, C., 2009. Measuring conditional cooperation: a replication study in russia. Experimental Economics 12, 87–92.

Kawagoe, T., Takizawa, H., 2012. Level-k analysis of experimental centipede games. Journal of Economic Behavior & Organization 82, 548–566.

Kessler, J. B., Leider, S., 2012. Norms and contracting. Management Science 58, 62–77.

Kimbrough, E. O., Robalino, N., Robson, A. J., 2014. The evolution of "theory of mind": Theory and experiments. Working Paper.

Kimbrough, E. O., Vostroknutov, A., 2015. Norms make preferences social. Journal of the European Economic Association, 1–31.

Kirkland, R. A., Peterson, E., Baker, C. A., Miller, S., Pulos, S., 2013. Meta-analysis reveals adult female superiority in "reading the mind in the eyes test". North American Journal of Psychology 15 (1), 121–146.

Kurzban, R., Houser, D., 2005. Experiments investigating cooperative types in humans: a complement to evolutionary theory and simulations. Proceedings of the National Academy of Sciences of the United States of America 102 (5), 1803–1807.

Lyons, M., Caldwell, T., Shultz, S., 2010. Mind-reading and manipulation- is machiavellianism related to theory of mind? Journal of Evolutionary Psychology 8 (3), 261–274.

Maestripieri, D., 2007. Machiavellian Intelligence: How Rhesus Macaques and HUmans Have Conquered the World. Chicago University Press.

Martino, B. D., O'Doherty, J. P., Ray, D., Bossaerts, P., Camerer, C., 2013. In the mind of the market: Theory of mind biases value computation during financial bubles. Neuron 80 (4), 1222–1231.

McCabe, K., Houser, D., Ryan, L., Smith, V., Trouard, T., 2001. A functional imaging study of cooperation in two-person reciprocal exchange. Proceedings of the National Academy of Sciences of the United States of America 98 (20), 11832–11835.

McCabe, K. A., Smith, V. L., LePore, M., 2000. Intentionality detection and "mindreading": Why does game form matter? Proceedings of the National Academy of Sciences of the United States of America 97, 4409–4409.

Nowak, M. A., Sigmund, K., 2005. Evolution of indirect reciprocity. Nature 437, 1291–1298.

Paal, T., Bereczkei, T., 2007. Adult theory of mind, cooperation, machiavelliansim: the effeffect of mindreading on social relations. Personality and Individual Differences 43 (3), 541–551.

Premack, D., Woodruff, G., 1978. Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences 1 (4), 515–526.

Preston, S. D., de Waal, F. B., 2002. Empathy: Its ultimate and proximate bases. Behavioral and Brain Sciences 25, 1–72.

Ridinger, G., 2015. Intentions versus outcomes: cooperation and fairness in a sequential prisoner's dilemma with nature. Working Paper.

Ridinger, G., 2016. Emotions, rule-following, and bargaining. Working Paper.

Ridinger, G., McBride, M., 2015. Money affects theory of mind differently by gender. PLOS ONE 10, e0143973.

Robalino, N., Robson, A., 2012. The economic approach to 'theory of mind'. Philosophical Transactions of the Royal Society of London B Biological Science 367 (1599), 2224–2233.

Robalino, N., Robson, A. J., In Press. The evolution of strategic sophistication. American Economic Review.

Rustagi, D., Engel, S., Kosfeld, M., 2010. Conditional coopeation and cocost monitorying explain success in forest ccommon management. Science 330, 961–965.

Seyfarth, R. M., Cheney, D. L., 2013. Affiliation, empathy, and the origins of theory of mind. Proceedings of the National Academy of Sciences of the United States of America 110, 10349–10356.

Singer, T., Fehr, E., 2005. The neuroeconomics of mind reading and empathy. American Economic Review 95, 340–345.

Stahl, D. O., Wilson, P. W., 1994. Experimental evidence on players' models of other players. Journal of Economic Behavior & Organization 25 (3), 309–327.

Sterelny, K., 2012. The Evolved Apprentice. Jean Nicod Lectures. MIT Press.

Sutton, J., Smith, P., Swettenham, J., 2010. Social cognition and bullying: Social inadequacy or skilled manipulation? Developmental Psychology 17 (3), 435–450.

Tomasello, M., 2014. The ultra-social animal. European Journal of Social Psychology 44, 187–194.

Toplak, M. E., West, R. F., Stanovich, K. E., 2011. The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. Memory and Cognition 39, 1275–1289.

Torralva, T., Kipps, C. M., Hodges, J. R., Clark, L., Bekinschtein, T., Roca, M., maria Lujan Calcagno, Manes, F., 2007. The relationship between affaffect decision-making and theory of mind in the frontal variant of fronto-temporal dementia. Neuropsychologia 45 (2), 342–349.

Vollm, B. A., Taylor, A. N., Richardson, P., Corcoran, R., Stirling, J., McKie, S., Deakin, J. F., Elliot, R., 2006. Neuronal correlates of theory of mind and empathy: A functional magnetic resonance imaging study in a nonverbal task. NeuroImage 29 (1), 90–98.

Whiten, A., Byrne, R., 1997. Machiavellian Intelligence II: Extensions and Evaluations. Machiavellian intelligence. Cambridge University Press.