Theory-of-mind Ability and Cooperation [*]

Garret Ridinger [†]
University of Nevada, Reno

Michael McBride [‡]
University of California, Irvine

September 2017

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.

[*] We thank the Experimental Social Science Laboratory at UC Irvine for use of its facilities. We also acknowledge comments and suggestions from seminar and conference participants. 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).

1 Introduction

It is widely held that human cooperation relies on the cognitive practice of attributing mental and emotional states to others. This capability—called theory of mind (ToM)—enables an individual to "mind read" another's desires, intentions, goals, and beliefs, thereby more accurately predicting another's behavior. A high ToM individual is thus more likely to identify and avoid cheaters but also recognize and successfully collaborate with mutual cooperators. Evidence for the ToM-cooperation connection comes from various sources. Evolutionary anthropologists find higher rates of cooperation among high-ToM humans than among other primates with lower ToM capacity (Stevens and Hauser, 2004; Warneken and Tomasello, 2006; Melis and Semmann, 2010).[1] Developmental psychologists find that children's cooperativeness increases as their ToM develops (Sally and Hill, 2006; Guroglu et al., 2009; Takagishi et al., 2010). Economic theories of cooperation through repeated interaction require individuals to hold beliefs about others' punishment strategies (Kreps et al., 1982; Fundenberg and Maskin, 1986; Abreu, 1988; Bo and Frechette, 2011), and more recent research by economists on social preferences identifies beliefs about others' beliefs, intentions, and utility functions as prominent factors in social cooperation (Fehr and Schmidt, 1999; Bolton and Okenfels, 2000; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006; Bicchieri, 2006; Kessler and Leider, 2012; Kimbrough and Vostroknutov, 2015).

Despite wide agreement that human cooperation depends on ToM, various gaps in our understanding remain. ToM is a complex cognitive phenomenon still only partially understood (Apperly, 2011). It develops significantly by the age of five, yet older children and adults still manifest differences in what might be called ToM ability, i.e., the accuracy and success of using ToM in a given social setting (Baron-Cohen et al., 1997, 1999, 2001; Ridinger and McBride, 2015). For example, females manifest higher ToM ability than adult males on average

[1] The ToM ability spectrum extends to non-human primates who appear to be unable to form higher order beliefs about mental states but can recognize intentional and goal-directed action (Premack and Woodruff, 1978; Cheney and Seyfarth, 2007; Seyfarth and Cheney, 2013) ToM is thus best understood as a set of cognitive faculties that includes more rudimentary skills that humans share with apes and monkeys and more advanced skills that are uniquely but universally human.

(Hall et al., 2000; Carroll and Chiew, 2006; Hoffman et al., 2010; Kirkland et al., 2013), and individuals with autism spectrum disorder manifest diminished ToM ability (Baron-Cohen et al., 1997, 1999, 2001).[2] A first gap is that no study to date has examined how differences in ToM ability in an adult human population correspond to differences in cooperativeness, even though the large majority of cooperative endeavors are undertaken by individuals with such differences.

A second gap is that the channel that links ToM ability and cooperativeness has not been conclusively identified. The first channel is that high-ToM-ability individuals form more accurate beliefs about others' cooperativeness and thereby find cooperators and avoid defectors. This theoretical link is widely accepted, yet we are the first to provide empirical evidence of this channel. A second possibility is that ToM fosters empathy (Preston and de Waal, 2002; Singer and Fehr, 2005; Vollm et al., 2006), so that high-ToM-ability individuals may be more empathetic and more cooperative, and a third possibility is that ToM encourages social manipulation (Whiten and Byrne, 1997; Maestripieri, 2007),[3] which suggests high-ToM-ability individuals may be less cooperative. Whether ToM ability influences cooperativeness solely via belief formation or also via fixed preference traits is unknown.

We use an experimental approach to identify the link between ToM ability and cooperation in an adult human population. We first present a model of norm-based utility to illustrate how variation in ToM ability can yield different rates of cooperation via beliefs or fixed preference trait channels. We then include a psychological measure of ToM in three experimental studies to identify the relationship between ToM ability and cooperation. The purpose is to identify conditions under which humans' ToM cognitive abilities are likely to produce cooperation, to determine the channels by which those cognitive abilities produce that cooperation, and to demonstrate if and how ToM ability proves to be advantageous.

[2] ToM abilty can also be experimentally manipulated (Kidd and Castano, 2013; Ridinger and McBride, 2015).

[3] Research examining the relationship between measures of Machiavellianism and ToM 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.

We report three main findings. First, whether higher ToM ability increases or decreases cooperativeness is contingent on multiple factors. In particular, (i) the population must include a high proportion of individuals who are willing to positively reciprocate, and (ii) there must also be a feature of the setting—e.g., sequencing of moves or signals of emotional states—which allows the high ToM individuals to leverage their cognitive capabilities. Though the former condition has been widely acknowledged,[4] we are the first to demonstrate how ToM ability determines the extent of cooperation given both conditions are present. Second, when ToM ability increases cooperativeness, it does so through improving the accuracy and precision of beliefs about others' behavior, not by a connection to fixed preference traits. Our results indicate that neither the empathy nor the "Machiavellian" argument is correct: ToM operates via beliefs not fixed traits. Third, by improving the ability to cooperate with reciprocators and defect on non-recipricators, high ToM ability yields highly significant payoff advantages. The payoff advantages are strongest in our experimental setting where direct inference of emotional states is possible. Evolutionary biologists and anthropologists have long believed that high ToM ability offered fitness advantages in humankind's distant past, thus creating selective pressures in favor of higher and higher ToM ability over evolutionary time (Flinn et al., 2005; Sterelny, 2012; Tomasello, 2014). Our experiment provides corroborating evidence from a laboratory setting for this evolutionary mechanism.

ToM has been called "the capstone attribute of human cognition" (Robalino and Robson, 2012), and experimental psychologists have developed multiple ways to measure it. We use the Reading the Mind in the Eyes Task (RMET) originally developed by Baron-Cohen et al. (2001). The subject is shown a cropped photo of an actor's eyes and then answers a multiple choice question about what the actor is thinking or feeling. The subject does this for thirty-six photos, and the number of correctly answered questions constitutes a measure of ToM. In effect, the RMET score measures the subjects' ability to accurately assess the emotional and cognitive state of others, a fundamental component of ToM abilty. The RMET has been extensively used in psychological studies of ToM and is widely held to be an accurate proxy for ToM ability (Baron-

[4] For example, see Nowak (2006); Nowak and Highfield (2011).

Cohen et al., 2001; Golan et al., 2006, 2007; Torralva et al., 2007). It's primary advantage for us is that it produces a rich distribution of RMET scores among a typical human population, unlike other binary measures of ToM that aim at assessing whether or not the individual (typically a child) has developed basic ToM (Baron-Cohen et al., 1985).

Aside from a few exceptions,[5] economists rarely use the term "theory of mind." Nonetheless, both macroeconomic and microeconomic theory incorporate ToM into 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 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' utility functions, others' behavior, others' beliefs about others' behavior, others' beliefs about others' beliefs about others' behavior, *ad infinitum.* Other game-theoretic concepts, such as those used in evolutionary game theory, do not assume ToM, while others, such as Level-k theory, allow for heterogeneity in ToM ability. As mentioned earlier, ToM ability 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, and 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 findings contribute to this literature by showing that variation in ToM ability corresponds to variation in cooperativeness in some settings but not others and that it does so via the beliefs not a preference channel. Future theoretical work that seeks to account for the role of cognition in social cooperation should account for these findings.

Only 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;

[5] See Robalino and Robson (2012), Kimbrough and Vostroknutov (2015), Georganas et al. (2015), Ridinger and McBride (2015) and Robalino and Robson (In Press).

Ridinger, 2016a), and ours is the first to use the RMET to study the relationship between ToM ability and cooperation. A separate neuroscientific literature uses magnetic imaging to identify neural correlates of ToM. This work reveals that prefrontal regions of the brain are more active in social interactions, including the Prisoners Dilemma game used in our study, thus providing evidence that the PD game is one in which ToM is active (McCabe et al., 2001) but without correlating ToM ability with degrees of cooperativeness. There exists, however, much 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 those 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 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 settings) to elicit subjects' conditional strategies, but we supplement with first and second-order belief elicitation and the RMET measure of ToM ability. Doing so allows us to separately identify the role of beliefs and fixed preference traits in the decision to cooperate.

2 Model

2.1 Preferences

Norm-utility models assume that an individual's utility decreases when her behavior differs from what she believes she ought 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 other models (inequity aversion, intentions, etc.) can account for the presence of reciprocating behavior in PD games, they cannot account for more purely cooperative behaviors and have trouble explaining some behavioral patterns (Clark and Sefton, 2001; Ridinger, 2016b). For example, standard versions of these preferences cannot explain, in a sequential PD game, the second mover's choice to cooperate after the first mover is known to defect, a behavior that occurs with non-trivial regularity in experiments (Clark and Sefton, 2001).[6] A norm model can account for this behavior because 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. 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 preference traits) are objects of our investigation. A norm model thus allows us to obtain testable predictions about the effect of ToM on cooperation.

Consider the following norm-based utility framework for an extensive form game with perfect recall and players $I = \{1, 2, \ldots\}$. At any information set h, let $N(h) \subseteq S_i(h)$ denote the set of actions available to player $i \in I$ at information set h that are prescribed (i.e., acceptable) by the norm. Say that an action available at h, denoted $s_i(h)$, is consistent with the norm if $s_i(h) \in N(h)$. If $s_i(h) \notin N(h)$, then $s_i(h)$ violates the norm. Given terminal node z, define $H_i(z)$ to be the set of information sets on the path that led to z for which i controls the choice.

[6] Some patterns from repeated-interaction experiments can also be understood in a norm framework. Many subjects begin cooperating and then reduce their cooperation over time. This is consistent with a norm model in which subjects are revising down their belief about others's cooperation. Similarly, initial defectors may learn to cooperate if they observe others adhering to a cooperative norm. However, we do not claim that norm models explain all experimental subjects' behavior but merely that the norm framework is flexible enough to account for a wide variety of behavior.

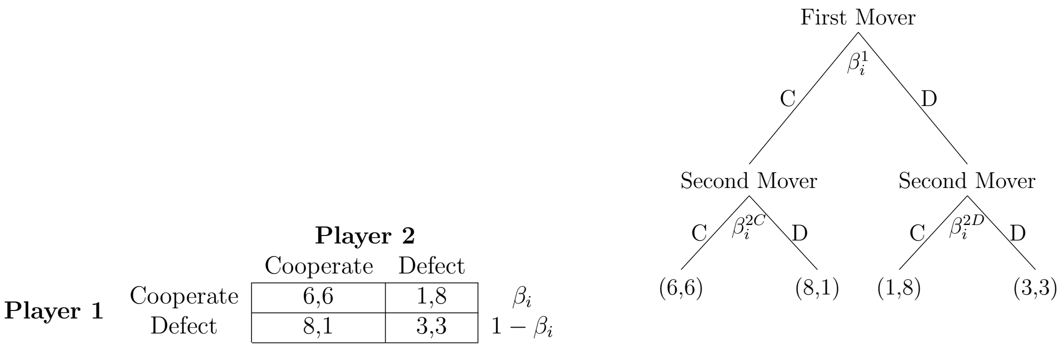
We now define i's utility function to be

$$u_i(z) = \pi_i(z) - \sum_{h \in H_i(z)} k_i \beta_i^h 1(s_i \notin N(h)),$$

where $\pi_i(z)$ is i's "material" payoff at terminal node z, k_i is i's exogenous norm sensitivity (or norm salience) that constitutes her fixed preference trait, $\beta_i^h \in [0,1]$ 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) at information set h, and $1(s_i \notin N(h))$ takes value 1 if $s_i(h)$ violates the norm and 0 otherwise. This specific functional form combines features of functions used in the literature. It is similar to that proposed by Lopez-Perez (2008), but, like (Bicchieri, 2006), has the penalty from norm violation increase in the proportion of community members that are believed to adhere to the norm.

Note that an information set corresponds to a particular role in the game, and the penalty from violating the norm associated with a particular role is only incurred if that role was on the path of play and the norm was violated in that role. A plan to violate a norm at an information set h incurs no penalty if h was never reached. This distinction is not important in simultaneous-move PD games because each player's single information set is on the equilibrium path. It will be relevant below in the sequential PD setting because the second mover could find herself in one of two possible roles, one role associated with the information set for after the first mover cooperates and another role associated with the information set for after the first mover defects.

We will assume that the norm for each role is cooperate. While this assumption may make intuitive sense for the simultaneous PD game, different norms seem reasonable for the two second-mover roles in the sequential PD game. For example, the norm for second movers in the sequential PD game could prescribe cooperate if the first mover cooperates but defect if the first mover defects, or the norm could prescribe cooperate if the first mover cooperates and do anything if the first mover defects. The model can accommodate the different norms, and our predictions will only differ with regard to behavior in the one second-mover role. To make predictions, however, we must choose our assumption about what norm is salient to the subjects. We here choose to assume that the norm in each role is to cooperate based primarily on evidence that a non-trivial minority of



		Player 2		
		Cooperate	Defect	
Player 1	Cooperate	6,6	1,8	β_i
	Defect	8,1	3,3	$1-\beta_i$

(a) Simultaneous

(b) Sequential

Figure 1: Simultaneous and Sequential Prisoner Dilemma Games

subjects in the second mover role cooperate even after the first mover defects, suggesting that such a norm is in operation for some subjects. However, we will revisit this question of the appropriate norm again when examining the experimental results. Moreover, we emphasize that we did not want influence the experimental subjects' perceptions of the appropriate norms, so we intentionally abstained from including language that would have signaled to the subjects that a particular norm is in effect.

2.2 Simultaneous Prisoners Dilemma

Figure 1(a) shows the simultaneous PD game we will use in our experiment. Payoff values represent the dollar amounts awarded in the experiment for different realized strategy profiles. As each player has only one information set in the simultaneous-move PD, we remove all information set subscripts. From a large community of agents, players 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

$$\begin{aligned} u_i\left(s_i=1,\beta_i\right) &= \beta_i\left(6-k_i\beta_i\left(0\right)\right)+\left(1-\beta_i\right)\left(1-k_i\beta_i\left(0\right)\right) \\ &= 1+5\beta_i, \end{aligned}$$

and her expected utility from deviating is

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

Cooperating is the best response for player i when

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

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_i). Cooperation depends on both the fixed preference trait k_i and beliefs β_i.

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_i^1 be player i's belief about the proportion of first movers that will cooperate. Let β_i^{2C} and β_i^{2D} 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 the second mover i's utility from cooperating is higher than her utility from defecting when

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

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

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

The first mover's expected utility from cooperating is

$$\begin{aligned} U_i\left(s_i=1, \beta_i^1, \beta_i^{2C}, \beta_i^{2D}\right) &= \beta_i^{2C}\left(6-k_i\beta_i^1\left(0\right)\right)+\left(1-\beta_i^{2C}\right)\left(1-k_i\beta_i^1\left(0\right)\right) \\ &= 1+5\beta_i^{2C}, \end{aligned}$$

and her expected utility from defecting is

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

Cooperating is player i's best response when

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

This more complicated expression for the first mover's decision is intuitive. When the actor believes others will cooperate (large β_i^1) and has a high norm sensitivity (large k_i), then she is more likely to cooperate. But her belief about what others do as second movers also affect her first-mover decision. If she believes that second movers are likely to cooperate if she cooperates (large β_i^{2C}), 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_i^{2D}), 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.

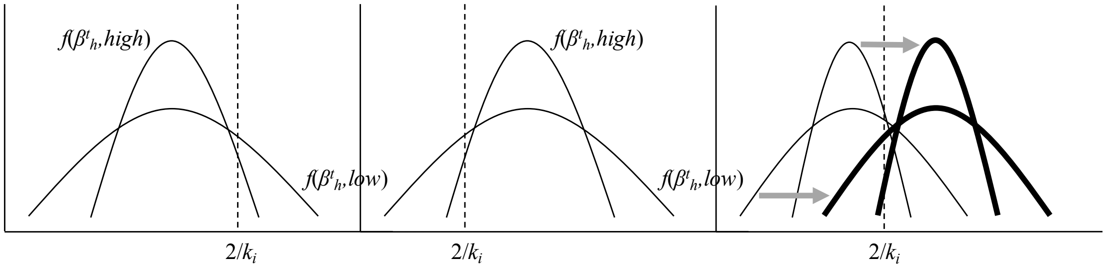
2.4 Theory of Mind Ability and Cooperation

A standard game-theoretic approach would fully specify an equilibrium. For example, a distribution of norm saliences in the population would be assumed, and each individual receives a noisy signal σ_i about that distribution, say its mean, where the distribution of errors in the signal is also known. After receiving her signal in the simultaneous PD game, the subject updates her belief about the likelihood that her partner will cooperate. A type-symmetric, cutoff-strategy Bayesian Nash Equilibrium could then be identified such that for each norm salience k_i there exists a $\sigma_i^* (k_i)$ such that the individual cooperates if $\sigma_i \geq \sigma_i^* (k_i)$ but defects otherwise. Calculation of $\sigma_i^* (k_i)$ would be non-trivial because each individual's posterior belief about the rate of cooperation must account for her belief about others' signals, the posterior beliefs that other agents will hold, and a belief that all actors agree on the equilibrium that is being played. However, this approach assumes extremely advanced cognition and ToM ability by all agents and thus violates the our premise that subjects vary in ToM ability.[7] We thus present a highly simplified, non-equilibrium account of ToM ability and cooperation, our purpose in this section being to merely illustrate how ToM can be associated with cooperation in subtle ways and then later identify specific predictions for the three experimental studies.

Suppose each player takes one of two ToM types: $ToM_i \in \{high, low\}$. Further suppose that each player's belief for an information set, β_i^h, is drawn from a distribution immediately prior to making a decision to cooperate or defect, with high ToM ability players drawing from p.d.f. $f\left(\beta_i^h, high\right)$ and low ToM ability players drawing from $f\left(\beta_i^h, low\right)$. Perhaps the high ToM ability players can use higher orders of thinking and low ToM ability players cannot (as in level k theory), or perhaps the high ToM ability agents receive more accurate signals about the distribution of norm saliences in the population. We are intentionally agnostic about the exact cognitive differences of these two types of individuals, but merely assume that there are differences that affect the formation of beliefs between the types that are fully captured by these belief distribution functions.

Consider the simultaneous PD setting, and suppose that the high ToM belief distribution has smaller variance but identical mean as the low ToM belief distribution. Further suppose that each individual (both high and low types) has a very small norm salience k_i so that $\frac{2}{k_i}$ is relatively large as depicted in Figure 2(a). Because the mean of each symmetric belief distribution is less than $\frac{2}{k_i}$ but the variance is smaller for the high types, there will be a higher proportion

[7] Experimental studies have found that non-equilibrium models can out predict equilibrium predictions in one-shot settings such as those in our experiments (Camerer et al., 2004; Crawford et al., 2013).



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

Figure 2: Examples of Distributional Predictions given Beliefs and ToM

of low types that obtain a belief draw above $\frac{2}{k_i}$ than high types that receive a belief draw $\frac{2}{k_i}$. Thus, with sharper precision for high types, the high types will cooperate at lower rates, on average, than low types. Figure 2(b) shows the opposite to be true if k_i is relatively large; when $\frac{2}{k_i}$ is relatively small, then the high types will be more likely to have a belief draw that is above $\frac{2}{k_i}$.

A few observations follow directly from this example. One is that differences in behavior between the high and low ToM ability individuals can arise without any differences in norm saliences. Additionally, whether the high ToM ability individuals cooperate at higher rates or at lower rates than the low ToM ability individuals depends on the norm saliences. Furthermore, if all individuals have identical beliefs, then any difference in cooperation rates between high and low types would be due to differences in norm salience.

This simple demonstration motivates a key feature of our experimental design. If we econometrically control for both beliefs and ToM ability in a regression with cooperation as the dependent variable, then the coefficient on ToM ability would pick up the correlation between ToM ability and unobserved norm saliences. If the coefficient is close to zero, then norm saliences are uncorrelated with ToM ability; if it is positive (negative), then high (low) ToM ability individuals have higher norm saliences. We could similarly control for norm saliences instead of beliefs, but measuring norm saliences are trickier to obtain.[8] Beliefs, on the other hand can be directly elicited via different well-studied procedures. Our experiment will thus elicit beliefs rather than norm saliences.

[8] The recently constructed rule-following task of Kimbrough and Vostroknutov (2015), which is intended to measure an individual's innate tendency to follow rules (norms), does not separately identify the norm salience from beliefs.

3 Study 1: Simultaneous PD

We now present the first of our three experimental studies. Our purpose in doing three is to systematically and exogenously adjust the interaction to allow increasing scope for high ToM ability individuals to exploit their cognitive advantage. The simultaneous PD game of Study 1 is meant to capture the simplest setting in which high ToM agents have the least ability to leverage their advantage.

3.1 Study 1 Procedures

A total of 112 subjects participated in Study 1. 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 was sent to a randomly-selected subset of the subject pool notifying them of the experiment and providing an 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 in more than one of our studies, 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 decisions made during the study. The experimental data are available from the authors upon request. This project was approved by the university's 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 the three parts of the experimental procedure. All decisions and responses are made through the mouse or keyboard. Each study lasted about one hour, divided into three parts.

Part I of this study is the thirty-six question Reading the Mind in the Eyes Task (Baron-Cohen et al., 2001). For each question, the subject is shown a cropped photograph of the eyes of an actor and a list of four emotions below the photograph; an example is shown in Figure 3. The subject is asked to select

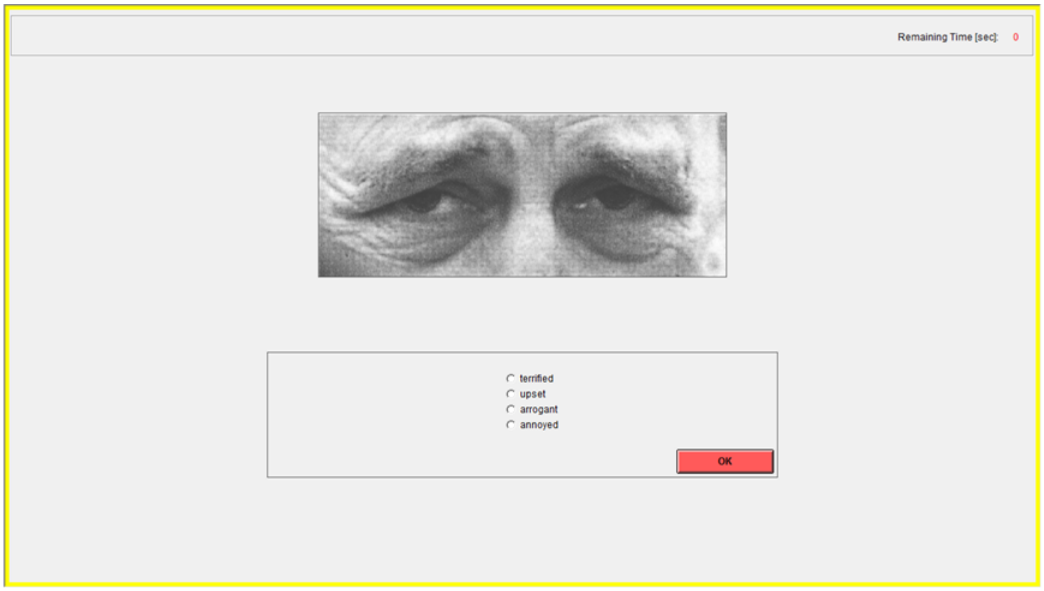


Figure 3: Example RMET Question

which of the four emotions best matches what the person in the photograph is thinking or feeling. 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 subject uses the mouse to select one of the four emotions on the screen. After making a selection, the screen goes blank for a brief moment and then advances to the next pair of eyes and four emotions so that each question is asked on its own screen with no feedback in between photos. 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 thirty-six photos. Subjects were not told how many photographs they would see, although they were told that the entire experiment would not last more than 1.5 hours.[9]

[9] 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.

Part II is the simultaneous PD game shown in Figure 1(b). The payoffs in the figure correspond the actual dollar values used. Subjects were randomly matched with another participant and asked to choose A or B. Option A is equivalent to cooperating and B equivalent to defecting, but those labels were not used to avoid unwanted priming of subjects. Subject were not informed of their partner's choice prior to making their decisions. After indicating their choices for the PD game but before learning the results, subjects reported their beliefs about others' actions and beliefs. Beliefs were elicited 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?"[10]

To incentivize belief reporting, we follow Charness and Dufwenberg (2006) and pay the subjects if her stated belief is within 5 percentage points of the correct belief. If a subject stated a belief within five percentage points of the truth, then she received $1. The payoff stakes are much lower than those for the PD game itself, thus reducing any incentive to hedge. 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 a belief less than 0.05 or greater 0.95. To measure accuracy of i's first-order belief, we define the average choice of others as,

$$\bar{s}_i^h = \frac{1}{i-1} \sum_{j \in N/\{i\}} s_j^h,$$

where N is the set of subjects in the session. We calculate the first-order belief accuracy for i as $|\beta_i^h - \bar{s}_i^h|$.[11] After reporting beliefs, the computer reveals the outcomes of the choices. A subject's payoff for Part II of the experiment is the

[10] We had to choose whether subjects do the RMET before PD or PD before RMET. Note that the RMET consisted of thirty-six questions of similar type, while the PD involves fewer questions of larger variety (choice and belief). We chose the first option primarily because we wanted subjects to remain engaged for the duration of the RMET, and placing it first ensured that the subjects were mentally fresh. We supposed that if the subjects' engagement declined as the RMET progressed, then it we be reinvigorated once the much shorted PD game started. We note that we are unaware of any studies that show evidence of spillover from the RMET to other decision tasks, but we kept the RMET first for all three studies so that any spillover would be similar across studies.

[11] We can represent the elicited second-order beliefs with additional notation. For each subject i, we define the second order belief as γ_i^h where h information set as before. Define

total of her payoff from the PD game and her payoff from the belief elicitation.

Part III is a questionnaire. Subjects were asked to report age, sex, and native language. This last question is important because RMET scores have been shown to depend partly on ethnic background not due to underlying ToM ability but because individuals of a non-American ethnicity have a disadvantage on the RMET due to less familiarity with the expressions of American faces (Adams et al., 2010). Subjects also completed the Cognitive Reflection Task (CRT) (Frederick, 2005).[12] The CRT is correlated with IQ and other mental heuristics (Toplak et al., 2011), and it has been found to be correlated with the RMET score (Baker et al., 2014). We control for both ethnicity and cognitive ability in our regressions (Ridinger and McBride, 2015).

3.2 Study 1 Predictions

As mentioned earlier, some have speculated that norm saliences should be positively correlated with ToM ability (Singer and Fehr, 2005) while others have suggested the opposite (Whiten and Byrne, 1997). We are not aware of any conclusive evidence that ToM ability is correlated with norm salience. We are aware of only one study that measures ToM ability and the propensity to follow norms, but we consider the results inconclusive.[13] With contradictory conjectures and no conclusive evidence, our null hypothesis is a zero correlation between ToM ability and norm saliences.

the average second order belief of others as

$$\overline{\beta}_i^h = \frac{1}{i-1} \sum_{j \in N/\{i\}} \beta_j^h,$$

which gives our measure of the accuracy of the second order beliefs as: $|\gamma_i^h - \bar{\beta}_i^h|$.

[12] The CRT consists of three questions. (1) A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? (2) If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? (3) In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?

[13] Ridinger (2016a) finds a positive and weakly significant relationship between the decision to follow a costly rule and the RMET, but because beliefs are not elicited, the study cannot distinguish whether that correlation is due to an association between ToM ability and norm saliences, ToM ability and beliefs, or both.

Study 1 is intended as a setting in which high ToM ability individuals cannot use their high ToM to their advantage. As a baseline prediction, we thus expect that belief accuracy will not be correlated with ToM ability. However, we do expect that subjects that predict more cooperation will be more likely to cooperate, consistent with the norm utility model. Overall, with no expected difference in norm saliences or beliefs between those with high ToM ability and low ToM ability, we should find no correlation between ToM ability and cooperation in Study 1.

Hypothesis 1 *In the Simultaneous PD Game in Study 1:*

(a) The likelihood of cooperation will not be correlated with RMET score.

(b) Belief accuracy will not be correlated with RMET score.

(c) The likelihood of cooperation will be positively correlated with belief in others' cooperation.

(d) The likelihood of cooperation will not be correlated with RMET score after controlling for beliefs.

3.3 Study 1 Results

Table 1 provides summary information for Study 1 and the other studies. Figure 4 plots the distribution of RMET scores for the different studies. Study 1 subjects' RMET scores range from 17 to 34, with a mean of 27.02, a median of 27, and a standard deviation of 3.88. These RMET scores are similar to those in studies 2 and 3 and to those from other studies (Ridinger and McBride, 2015).[14]

The claims in Hypothesis 1 are confirmed. Figure 5 shows the raw rates of cooperation when dividing the sample into those with below-median RMET scores (low) and above-median RMET scores (high). The rates of cooperation are nearly identical and not statistically different. Table 2 shows the results from four regressions.[15] Regression (1) finds that, consistent with Hypotheses 1(a), there is no overall correlation between RMET score and the likelihood of

[14] No statistical differences in RMET scores between treatments were found using a one-way ANOVA and a post-hoc Tukey test. See the Appendix for details.

[15] Reported regressions use Ordinary Least Squares (OLS) which could be a potential issue when the dependent variable is binary. We conducted the analysis using both Probit and Logit specifications as robustness checks and the results are similar. Due to this, we report the OLS results for ease of presentation.

Table 1: Summary Statistics

	Overall	Simultaneous PD	Sequential PD	Eyes BC	Eyes PD
	mean (sd)	mean (sd)	mean (sd)	mean (sd)	mean (sd)
RMET	27.21	27.02	27.51	27.77	26.43
	(3.61)	(3.88)	(3.32)	(3.47)	(3.76)
Female	0.58	0.68	0.50	0.52	0.61
	(0.49)	(0.47)	(0.50)	(0.50)	(0.48)
Age	20.22	19.88	20.31	20.33	20.49
	(1.62)	(1.28)	(2.00)	(1.53)	(1.69)
Native English Speaker	0.52	0.49	0.51	0.57	0.53
	(0.50)	(0.50)	(0.50)	(0.49)	(0.50)
CRT	1.10	1.19	1.16	1.07	0.88
	(1.14)	(1.20)	(1.12)	(1.19)	(1.02)
Number of Economic Courses	1.41	1.42	1.66	1.43	1.00
	(2.29)	(2.40)	(2.86)	(1.78)	(1.80)
Number of Statistics Courses	1.14	1.05	1.14	1.43	1.00
	(2.20)	(2.40)	(2.86)	(1.78)	(1.24)
Take Home Pay	13.02	12.32	12.12	15.89	12.35
	(3.28)	(3.91)	(2.55)	(3.28)	(3.40)
Observations	363	112	106	77	68

cooperating. Regression (2) finds that having a higher RMET score is associated with lower prediction of others's cooperation, though this difference is only moderately significant. Regression (3) shows that having a higher RMET score is not associated with higher belief precision.[16] Together these regressions indicate that ToM ability does not play a strong role in belief formation in the simultaneous PD setting, as indicated by Hypothesis 1(b). Regression (4) adds beliefs as an independent variable to regression (1) and shows that subjects who believe in higher rates of others' cooperation will be more likely to cooperate, consistent with our norm-utility model. However, there is still no correlation between RMET score and the likelihood to cooperate, indicating that there is no correlation between RMET score and the unobserved norm salience. These findings match the Hypothesis 1(c) and (d) predictions. As expected, this simultaneous PD setting did not allow the high ToM individuals to leverage their cognitive advantage.

[16] Regressing RMET on second-order beliefs shows that the regression coefficient for RMET is not significant in predicting second-order beliefs. Additionally, higher RMET is not associated with increased second-order belief precision. For additional details, please see Appendix.

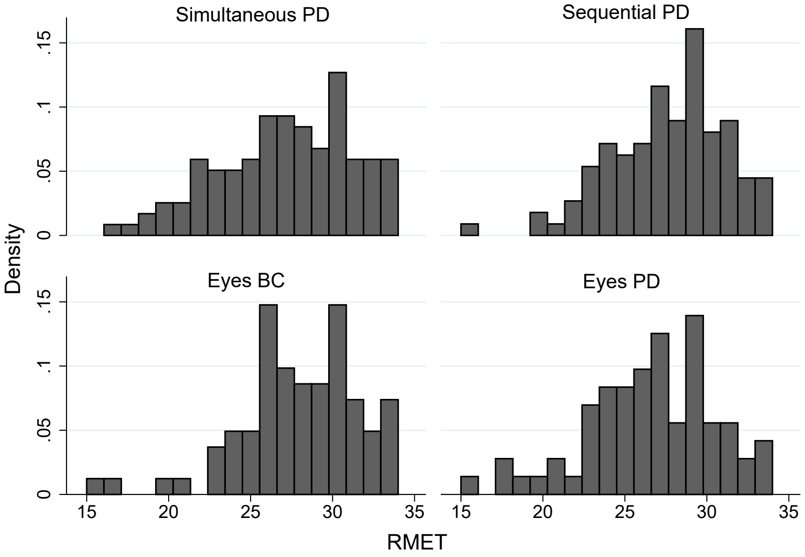


Figure 4: Histogram of RMET Scores for Simultaneous PD

4 Study 2 Sequential PD

4.1 Study 2 Procedures

104 subjects participated in Study 2, which is identical to Study 1 except in Part II the subjects play the sequential PD game from Figure 1(b) instead of the simultaneous PD game. The strategy method is used to obtain each subject's full strategy. Subjects were randomly and anonymously paired with another participant and explained that they will report their plan of what to do if selected as first mover and the plan of what to do as second mover conditional on the possible actions by the first mover. That is, the subject selects A or B as the choice for when the first mover, A or B as the choice for when the second mover after the first mover chose A, and A or B as the choice for when the second mover after the first mover chose B. It is also explained that, after selecting their plans, the computer will randomly assign the subjects to the two roles and carry out the plans selected by the subjects.

After selecting their plans, the subjects are asked to report their beliefs about the plans selected by the other subjects. Specifically, the screen displays

Table 2: Predicting Cooperation by RMET and Belief in Simultaneous PD

	(1)	(2)	(3)	(4)
	Cooperate	β_i	$\lvert\beta_i - \bar{s}_i\rvert$	Cooperate
RMET	-0.01	-0.01*	-0.00	0.00
	(0.01)	(0.01)	(0.00)	(0.01)
β_i				0.96***
				(0.14)
Female	0.31***	0.12**	-0.00	0.20**
	(0.10)	(0.06)	(0.04)	(0.10)
Age	-0.05	0.04*	0.00	-0.08***
	(0.04)	(0.02)	(0.01)	(0.03)
Native English Speaker	-0.08	-0.01	0.02	-0.07
	(0.10)	(0.05)	(0.03)	(0.09)
CRT	0.01	-0.02	0.01	0.02
	(0.04)	(0.02)	(0.01)	(0.04)
Intercept	1.54**	0.15	0.24	1.40**
	(0.74)	(0.44)	(0.25)	(0.55)
N	112	112	112	112
R^2	0.136	0.118	0.032	0.358

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

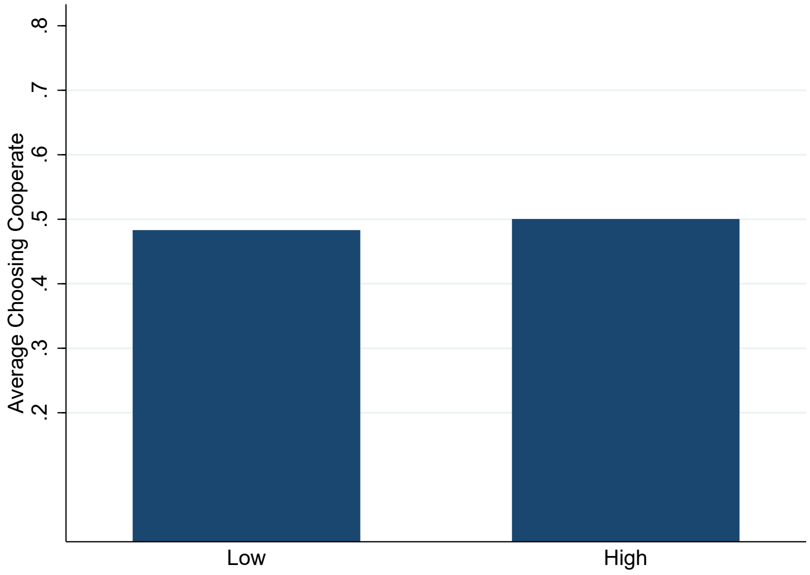


Figure 5: Average Cooperation by Low and High RMET Scores in Simultaneous PD

the following: "When asked to select their plan of what to do as First Mover: (1) What percentage of other people in the room choose A? (2) What is the average answer of the other people in the room to Question (1) above?" Similar questions are asked about what to do as second mover if the first mover cooperated and what to do as second mover if the first mover defects.

After reporting beliefs, the roles were randomly determined by the computer, and the computer carried out the indicated plans on the subjects' behalf to end Part II. As with Study 1, a subject was paid based on the outcome of the PD game and the belief elicitation.

4.2 Study 2 Predictions

Prior studies of the sequential PD game have found that typically 30-50% of the subjects in the second mover role choose the reciprocating strategy of cooperating after the first mover cooperates and defecting after the first mover defects (Clark and Sefton, 2001; Ahn et al., 2007; Dhaene and Bouckaert, 2010). A second mover that expects others to be playing this strategy should thus believe that there will be much higher rates of cooperation after first mover cooperation than after first mover defection. However, we expect that not all players will equally anticipate second-mover reciprocity. The experiment itself provides no hints, subtle or otherwise, that reciprocating strategies may be used. Anticipating reciprocal behavior by other second movers requires a subject to put herself into the mind of the other subjects to anticipate their responses, thus suggesting a possible advantage for high ToM ability individuals. We predict that high ToM ability subjects will expect more reciprocity by second movers as compared to low ToM ability subjects. Moreover, with these different beliefs, high ToM ability subjects should themselves be more likely to cooperate after cooperation and defect after defecting than low ToM ability subjects.

As shown in Section 2.4, the anticipated second-mover behavior affects the decision of a first mover. If high ToM ability subjects anticipate more reciprocity, then as first movers they should be more willing to cooperate to induce positive reciprocity by second movers. Moreover, extending the logic, they should also expect more first-mover cooperation, providing another reason to cooperate as first mover as they expect more of the other first movers to comply with the norm to cooperate. We thus expect high ToM ability individuals to cooperate at higher rates as first movers than low ToM ability subjects.

Based on our confirmation of Hypothesis 1(c), we expect subjects' rates of cooperation at each decision node in the game tree will increase in their belief that others cooperate at that node. Finally, based on our confirmation of Hypothesis 1(d), we expect that once we control for beliefs, there will be no added correlation between ToM ability and rates of cooperation at any node in the game tree.

Hypothesis 2 *In the Sequential PD Game in Study 2:*

(a) The likelihood of cooperation will be positively correlated with RMET score for first movers and second movers after cooperation but negatively correlated with second mover cooperation after defection.

(b) Belief accuracy will be correlated with RMET score.

(c) The likelihood of cooperation will be positive correlated with belief in others' cooperation at each decision node.

(d) The likelihood of cooperation will not be correlated with RMET score at any decision node after controlling for beliefs.

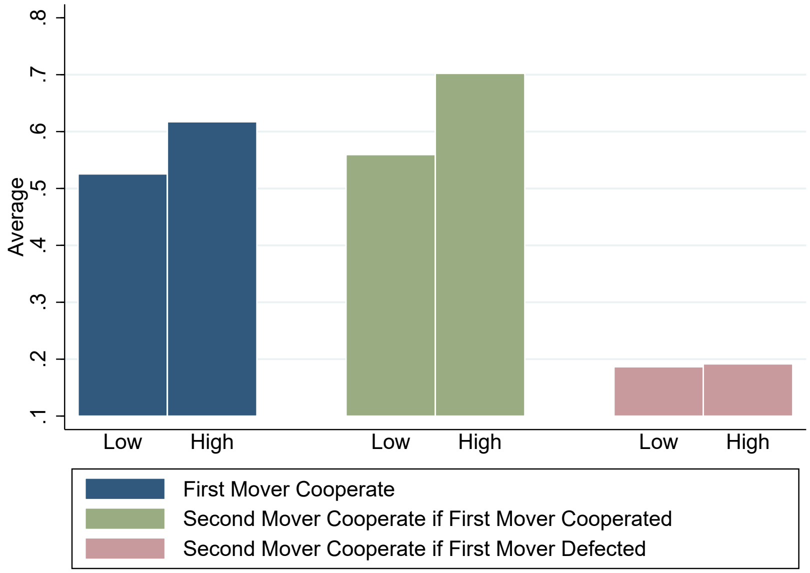


Figure 6: Average Cooperation by Low and High RMET Scores in Sequential PD

4.3 Study 2 Results

Evidence supports most but not all of the claims in Hypothesis 2. Figure 6 shows the percent of subjects choosing cooperate by information set for the high and low RMET score subjects. As predicted by Hypothesis 2(a), the high RMET subjects cooperate at a higher rate as first movers and as second movers after first mover cooperation. The difference in cooperation rates is statistically significant for both cases. Contrary to Hypothesis 2(a), the rates of cooperation between high and low RMET subjects are identical when acting as second mover after defection. High RMET subjects are not more likely to cooperate than low RMET subjects should the first mover defect.

Table 3 reports the results of six regressions, i.e., two regressions (one for mean belief and one for belief accuracy) for each of the three information sets at which a belief was reported. Regressions (1)-(2) reveal that RMET score is positively correlated with the mean belief when first mover and second mover after cooperation; at those two information sets, the higher one's RMET, the higher percentage of others one believes will cooperate. Mean belief is not correlated with RMET when second mover after defection. Regressions (4)-(5) show that RMET score is positively correlated with belief accuracy when first mover and second mover after cooperation but not when second mover after defection.[17] These results match the prediction in Hypothesis 2(b) for first movers and second movers after cooperation, but not for second movers after defection.

Table 4 reports the results of six regressions predicting cooperation rates. Regressions (1)-(3) match what is shown in Figure 6, confirming that RMET score is positively correlated with the likelihood of cooperation when first mover and second mover after cooperation, but not when second mover after defection. Two of the three cases match the prediction in Hypothesis 2(a). Regressions (4)-(6) control for beliefs. As predicted in Hypothesis 2(c), beliefs strongly predict cooperation, as evidenced by the positive and highly significant coefficients on beliefs in regressions (4)-(6). We also see that the positive coefficients on RMET score in regressions (1) and (2) now go to zero once beliefs are controlled. As predicted by Hypothesis 2(d), the effect of RMET on cooperation found in regressions (1) and (2) is fully accounted for once beliefs are controlled.

ToM ability is not correlated with cooperation in the simultaneous PD game, but it is in the sequential PD game, suggesting that the role that ToM plays in fostering cooperation depends on a confluence of factors. Chief among them is the presence of reciprocators that individuals with high ToM ability accurately predict are present. Moreover, when ToM does influence cooperation, our analysis reveals that it does so through beliefs and not fixed preference traits. Once beliefs are controlled, there is no additional influence of ToM ability on cooperation. That ToM ability is not correlated with beliefs or behavior of second movers after defection could be due to multiple factors. There could be lack of agreement on what is the appropriate behavior after first mover defection. Cooperating is an altruistic act, but defecting could also be considered an appropriate way to punish a defecting first mover. A lack of agreement on what is appropriate after a defection could generate additional noise in the data that swamps out any effect of having high ToM ability. Another possibility is that defecting may be an obvious reply to first-mover defection for all subjects, in which case increased ToM ability within our sample would not lead to any difference among second movers after defection.

[17] Similar results are found for second-order beliefs and accuracy. For additional details, please see Appendix.

Table 3: Predicting First Order Beliefs and Accuracy by RMET Score in Sequential PD

	(1) β_i^1	(2) β_i^{2C}	(3) β_i^{2D}	(4) $\lvert\beta_i^1 - \bar{s}_i^1\rvert$	(5) $\lvert\beta_i^{2C} - \bar{s}_i^{2C}\rvert$	(6) $\lvert\beta_i^{2D} - \bar{s}_i^{2D}\rvert$
RMET	0.02***	0.02**	0.01	-0.01**	-0.01**	0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Female	0.07	0.02	-0.04	-0.05	-0.04	-0.06
	(0.06)	(0.07)	(0.07)	(0.04)	(0.05)	(0.05)
Age	0.02**	-0.01	-0.02*	-0.00	0.01	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Native English Speaker	-0.00	-0.08	-0.15**	0.02	-0.02	-0.08*
	(0.06)	(0.07)	(0.07)	(0.04)	(0.04)	(0.05)
CRT	-0.06**	0.00	-0.02	0.01	-0.01	-0.00
	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)
Intercept	-0.38	0.17	0.65*	0.65**	0.50*	0.53**
	(0.31)	(0.35)	(0.39)	(0.28)	(0.27)	(0.26)
N	104	104	104	104	104	104
R^2	0.148	0.125	0.077	0.115	0.117	0.100

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

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

Table 4: Predicting Cooperation by RMET and Beliefs in Sequential PD

	(1) First Mover Cooperate	(2) Second Mover Cooperate if First Mover Cooperated	(3) Second Mover Cooperate if First Mover Defected	(4) First Mover Cooperate	(5) Second Mover Cooperate if First Mover Cooperated	(6) Second Mover Cooperate if First Mover Defected
RMET	0.03*	0.03**	0.01	0.00	0.02	0.01
	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
β_i^1				1.07***		
				(0.13)		
β_i^{2C}-β_i^{2D}				0.00**		
				(0.00)		
β_i^{2C}					0.53***	
					(0.16)	
β_i^{2D}						0.43***
						(0.10)
Female	0.11	-0.07	0.02	0.03	-0.08	0.04
	(0.11)	(0.10)	(0.07)	(0.08)	(0.09)	(0.07)
Age	0.04**	0.01	-0.03*	0.01	0.02	-0.02
	(0.02)	(0.01)	(0.02)	(0.02)	(0.02)	(0.01)
Native English Speaker	0.01	0.02	-0.19**	-0.00	0.07	-0.13*
	(0.10)	(0.10)	(0.08)	(0.08)	(0.09)	(0.08)
CRT	-0.05	-0.09**	-0.05*	0.00	-0.09**	-0.04*
	(0.05)	(0.04)	(0.03)	(0.04)	(0.04)	(0.03)
Intercept	-0.70	-0.36	0.76*	-0.21	-0.45	0.48
	(0.55)	(0.52)	(0.44)	(0.43)	(0.49)	(0.37)
N	104	104	104	104	104	104
R^2	0.102	0.113	0.226	0.481	0.213	0.356

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

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

5 Study 3: RMET Eyes Sequential PD

Our final study constructs a setting in which there is much larger scope for ToM ability to impact behavior by providing subjects with emotional signals about the behavior of their partners. We have two main purposes. First, we want to examine the range of behaviors that ToM ability fosters, i.e., to show that high ToM ability individuals may actually be less cooperative in some interactions and not just more cooperative as in Study 2. Second, by pairing subjects against the computer, we can remove other-regarding preference considerations. Note that Studies 1 and 2 establish that high ToM ability is not correlated with the strength of other-regarding preferences, so removing other-regarding considerations allows us to focus solely on the role and advantages of predictive accuracy.

This study consists of two separate conditions, A and B. Condition B, the primary condition of interest, has subjects take the role of first movers in a sequential PD game, observe each of the pairs of eyes from the RMET, and then choose whether to cooperate or defect as first mover with the eyes being a signal of the second mover's emotional state. The decision of the pair of eyes in the second-mover role is determined by the outcome of Condition A in which a separate set of subjects predict the second-mover behavior of the pair of eyes within a beauty contest game. We discuss this beauty contest game first.

5.1 Condition A: Eyes Beauty Contest

Condition A, conducted with 77 subjects, consists of two parts. Part I and III are the same as Studies 1 and 2. Part II is the Eyes Beauty Contest. In this part of the study, the subjects are shown the sequential PD game used in Study 2. Each subject is next told that she will be shown a photo of a pair of eyes and asked to select what she thinks someone with that pair of eyes would do as second mover in that sequential PD game. That is, the subject reports what she thinks someone with those eyes would do if the first mover cooperates and what someone with those eyes would do if the first mover defects. Before the subject reports her answers, she is told that she will provide these two answers for thirty-six photos, that at the end of the experiment one answer from one of the thirty-six photos will be selected for payment, and that if her answer for that photo matches what a majority of others in the room reported for that answer then she would receive \$5. She receives \$0 if her answer does not match the majority. The subject completes this task for each of the thirty-six photos from the RMET.

5.2 Condition B: Eyes PD

Condition B, conducted with 68 subjects, consists of three parts. Parts I and III are the RMET and questionnaire as done in Studies 1 and 2. Part II is the Eyes PD game. The subject is shown the first pair of eyes from the RMET and the sequential PD game from Condition A. She is then told that she is the first mover, that the person with the pair of eyes is the second mover, and that the second mover's decision is based on what a majority of people in a previous experiment thought a person with those eyes would do as second mover. The subject then selects A or B. On the same screen the subject is asked to predict what the majority of other people in the previous experiment though a person with this pair of eyes would do as second mover after the first mover cooperates and after the first mover defects. The subject repeats this for each of the thirty-six RMET photos. It is explained that one of the thirty-six photos will be selected at random for monetary payment. The subject's payment from the PD portion is calculated based on her action and on the second mover's decision as determined by the responses in the Condition A beauty contest. One of the belief questions is also selected for payment, with a correct answer receiving \$2, a fact explained before the subject is shown the first photo.

5.3 Study 3 Predictions

Given that RMET measures ToM ability by measuring the accuracy in reading emotions, we anticipate that subjects with high RMET subjects will respond more sharply to the emotional cues in the photos than the low RMET subjects because they will more accurately predict what the beauty contest majority predicted. Consequently, the high RMET subjects should be more likely to choose cooperate when the second mover eyes will cooperate after first mover cooperate, and be less likely to cooperate when the second mover eyes will defect after first mover cooperation. We further expect difference this to yield a significant payoff advantage to high RMET subjects because they will be more likely to initiate cooperative play with positive reciprocators and avoid the defection of non-recipricators.

Hypothesis 3 *In the Eyes PD Game in Study 3:*

(a) Subjects with high RMET in Condition B will more accurately predict the beauty contest majority from Condition A.

(b) RMET score should be positively correlated with first mover cooperation in Condition B for eyes in which the majority from Condition A reported would cooperate after first mover cooperation; but RMET score should be negatively correlated with first mover cooperation when the Condition A majority reported defection after first mover cooperation.

(c) RMET score should be positively correlated with expected payoffs.

5.4 Study 3 Results

The evidence largely supports the claims in Hypothesis 3. Table 5 regresses RMET on the number of times subjects from Condition B correctly predicted what the majority of subjects in Condition A predicted. The high RMET subjects from Condition B are better at predicting the Condition A beauty contest majority than the low RMET subjects, consistent with Hypothesis 3(a). The starkest difference is for those eyes that the beauty contest subjects overwhelmingly predict defect after cooperation. Interestingly, the high RMET subjects are particularly good at identifying when the second mover will not positively reciprocate after first mover cooperation.

Higher belief accuracy for high RMET subjects should yield different behavior as first movers. Figure 7 partitions the beauty contest responses into three categories: eyes very likely to defect after cooperation such that less than 35% of the beauty contest subjects predicting cooperate after first-mover cooperation, eyes whose behavior after first mover cooperation is unclear with between 35-65% predicting cooperate, and eyes very likely to cooperate after first-mover cooperation with more than 65% predicting cooperate. We see that the high

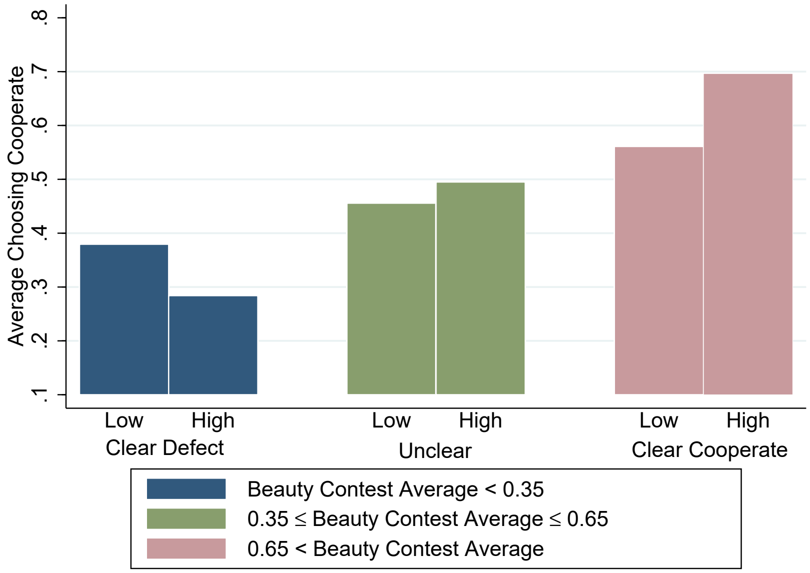


Figure 7: Average Cooperation by Low and High RMET Scores in Eyes PD

RMET first movers are much less likely to cooperate than the low RMET first movers when reciprocating cooperation is unlikely, and more likely to cooperate when reciprocating cooperation is likely, as predicted in Hypothesis 3(b). Additional evidence from regressions is reported in Table 6. RMET is not correlated with first mover cooperation in the Eyes PD condition as shown in regression (1), which is to be expected because the high RMET subjects might be more likely to cooperate against some eyes but less likely to cooperate against others. Regressions (2) shows that cooperation is more likely as the percent of beauty contest subjects predicting cooperation increases, indicating that the Eyes PD subjects are reacting to the emotional cues in the photos that prompted the beauty contest subjects' predictions. The key control in regression (3) is the interaction between RMET and beauty contest average; the positive and highly significant coefficient indicates that the higher one's RMET score, the stronger the behavioral response to the emotional cues in the photoes.

Table 5: Predicting Number of Correct Predictions of Beauty Contest Majority by RMET Score

	(1) Correct if First Mover Cooperated	(2) 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

Results are from Ordinary Least Squares regressions. Robust standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 6: Predicting Cooperation and Payoffs by RMET Score and Beliefs in Eyes PD

	(1) Cooperate	(2) Cooperate	(3) Cooperate	(4) Expected Payoff (Unclear Eyes)	(5) Expected Payoff (Clear Eyes)
RMET	0.00	0.00	-0.02***	0.00	0.02***
	(0.00)	(0.00)	(0.01)	(0.00)	(0.01)
Average Beauty Contest β_i^{2C}		0.23*** (0.05)	-1.04*** (0.36)		
RMET X Average Beauty Contest β_i^{2C}			0.05*** (0.01)		
β_i^{2C}		0.38***	0.38***		
		(0.02)	(0.02)		
β_i^{3D}		0.01	0.01		
		(0.02)	(0.02)		
Female	0.02	0.07***	0.07***	0.00	-0.01
	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)
Age	0.01**	0.01*	0.01*	-0.01	-0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Native English Speaker	-0.01	-0.02	-0.02	0.00	-0.01
	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)
CRT	-0.05***	-0.03***	-0.03***	0.03***	0.06***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
Intercept	0.11	-0.17	0.43*	3.81***	3.36***
	(0.15)	(0.14)	(0.22)	(0.15)	(0.35)
N	2448	2448	2448	1496	952
ρ	0.03***	0.00***	0.00***	0.53***	0.62***
Model χ^2	64.39	590.70	606.41	20.35	32.14

Regressions are from random effects linear regression with session fixed effects. Standard errors in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Regressions (4) and (5) regress expected earnings on RMET score. When including all eyes, RMET score does not correlate with expected earnings. Yet, when only the eyes with a strong prediction are included, RMET score is positively correlated with expected earnings as predicted by Hypothesis 3(c). Moreover, the impact of RMET score on payoff is a meaningful one. The average expected payoff for a picture of the unclear eyes is \$3.73 and an increase of one standard deviation in RMET increased the expected payoff by \$0.06. While increased RMET significantly increases payoff for eyes that are highly likely to cooperate, the largest percentage increase in payoffs occurs for subjects who are more likely to recognize the eyes highly likely to defect. The average expected payoff for picture of eyes thought likely to defect is \$3.20 and an increase of one standard deviation in RMET increases the expected payoff by \$0.07.

6 Discussion

Our study set out to understand how differences in ToM ability influence cooperation. We introduced a model framework that demonstrates how the effects of ToM ability on cooperation can vary. We presented three experimental studies, each with successively larger scope for high ToM ability subjects to use their ability to their advantage. Our first study of the simultaneous PD finds no correlation between ToM ability and the propensity to cooperate. Cooperation is driven largely by the likelihood with which the subject believes the partner will cooperate, but higher ToM ability subjects are no more accurate in their beliefs than lower ToM ability subjects in the simultaneous PD setting. There is also no correlation between ToM ability and a fixed preference trait that predisposes cooperation or defection. Our second study of the sequential PD game finds a correlation between ToM ability and the propensity to cooperate as first mover and as second mover after first-mover cooperation; that is, the higher the subject's ToM ability, the more likely the subject is to cooperate in those two roles. Further examination reveals that ToM ability is enabling more accurate beliefs about positive reciprocity and, again, not correlated with a fixed preference trait. Our third study of a PD game with emotional signals finds that the higher ToM ability subjects are much more accurate at predicting the second-mover responses as chosen by an independent set of subjects. Importantly, the high ToM ability subjects are particularly good at identifying when second-movers will not positively reciprocate, and having higher ToM ability leads to a statistically significant boost in expected payoff.

Our project yields three larger lessons about the role of ToM in fostering cooperation. First, whether higher ToM ability leads to more cooperation or less cooperation is contingent on multiple factors such as the presence of a high proportion of individuals who are willing to positively reciprocate and some feature of the setting which allows the high ToM individuals to leverage their skills in forming accurate beliefs about reciprocation. Second, when ToM ability does improve cooperation, it does so through improving the accuracy and precision of beliefs about others' behavior, not by a deeper association between ToM ability and fixed preferences traits. Psychologists find that increasing rates of cooperation in children coincide with ToM development, and our study extends our understanding to adults. Once basic ToM ability has developed normally, additional ToM ability might not foster norm following directly via preferences but acts as a cognitive skill in belief formation that can lead to norm adherence. Third, higher ToM ability can yield significant payoff advantages, a finding consistent with the belief that ToM ability is an evolved trait that offered fitness advantages in humankind's distant past.

Future theoretical and experimental work has many avenues to consider. There are, for example, a wide range of strategic scenarios in which ToM should matter, and future experiments should investigate in which of them ToM does indeed matter. Future theoretical work can justifiably assume that ToM ability operates through belief formation and not fixed preference traits. What is needed is a broad theory of the conditions under which variation in ToM ability yields differences in behavior. Our prediction that ToM ability would matter more in the sequential and PD eyes PD settings was based on a reasonable premise that ToM ability might foster better predictions when there are more roles must be envisioned. Additional theoretical and experimental work is needed to identify the deeper processes at work.

References

Abreu, D., 1988. On the theory of infinitely repeated games with discounting 56, 383–396.

Adams, R. B., Rule, N. O., Franklin, R. G., Wang, E., 2010. Cross-cultural reading the mind in the eyes: An fmri investigation 22, 97–108.

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.

Apperly, I., 2011. Mindreaders: The Cognitive Basis of "Theory of Mind". Psychology Press.

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., Jolliffe, T., Mortimore, C., Robertson, M., 1997. Another advanced test of theory of mind: Evidence from very high functioning adults with autism or asperger syndrome 38, 813–822.

Baron-Cohen, S., Leslie, A. M., Frith, U., 1985. Does the autistic child have a "theory of mind"? Cognition 21, 37–46.

Baron-Cohen, S., O'Riordan, M., Stone, V., Jones, R., Plaisted, K., 1999. Recognition of faux pas by nonormal developing children and children with asperger syndrome or high-functiong austism 29, 407–418.

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.

Bo, P. D., Frechette, G. R., 2011. The evolution of cooperation in infinitely repeated games: Experimental evidence 101, 411–429.

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.

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

Carroll, J. M., Chiew, K. Y., 2006. Sex and discipline differences in empathising, systemising and autistic symptomatology: evidence from a student population 36, 949–957.

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

Cheney, D. L., Seyfarth, R. M., 2007. Baboon Metaphysics: The Evolution of a Social Mind. University of Chicago Press.

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

Crawford, V. P., Costa-Gomes, M. A., Iriberri, N., 2013. Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications 51, 5–62.

Dhaene, G., Bouckaert, J., 2010. Sequential reciprocity in two-player, two-stage games: An experimental analysis. Games and Economic Behavior 70, 289–303.

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.

Fundenberg, D., Maskin, E., 1986. The folk theorem in repeated games with discounting or with incomplete information 54, 533–554.

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.

Guroglu, B., van den Bos, W., Crone, E. A., 2009. Fairness considerations: Increasing understanding of intentionality during adolescence 104, 398–409.

Hall, J. A., Carter, J., Horgan, T., 2000. Gender differences in the nonverbal communication of emotion. In: Fisher (Ed.), Gender and Emotion: Social Psychological Perspectives. Cambridge University Press, pp. 97–117.

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

Hoffman, H., Kessler, H., Eppel, T., Rukavina, S., Traue, H. C., 2010. Expression intensity, gender and facial emotion recognition; women recognize only subtle facial emotions better than men. 135, 278–283.

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.

Kidd, D. C., Castano, E., 2013. Reading literary fiction improves theory of mind. Science 342, 377–380.

Kimbrough, E. O., Vostroknutov, A., 2015. Norms make preferences social. Journal of the European Economic Association 14 (3), 608–638.

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.

Kreps, D. M., Milgrom, P., Roberts, J., Wilson, R., 1982. Rational cooperation in the finitely repeated prisoner's dilemma 27, 245–252.

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.

Lopez-Perez, R., 2008. Aversion to norm-breaking: a model. Games and Economic Behavior 64, 237–267.

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.

Melis, A. P., Semmann, D., 2010. How is human cooperation different? 365, 2663–2674.

Nowak, M. A., 2006. Five rules for the evolution of cooperation. Science 314, 1560–1563.

Nowak, M. A., Highfield, R., 2011. SuperCooperators: Altruism, evolution, and why we need each other to succeed. New York: Free Press.

Paal, T., Bereczkei, T., 2007. Adult theory of mind, cooperation, machiavelliansim: the efeffect 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., 2016a. Emotions, rule-following, and bargaining. Working Paper.

Ridinger, G., 2016b. Intentions versus outcomes: cooperation and fairness in a sequential prisoner's dilemma with nature. Working Paper Available at SSRN: https://ssrn.com/abstract=2841833.

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.

Sally, D., Hill, E., 2006. The development of interpersonal strategy: Austism, theory-of-mind, cooperation and fairness 27, 73–97.

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.

Stevens, J. R., Hauser, M. D., 2004. Why be nice? psychological cconstraint on the evolution of cooperation 8, 60–65.

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

Takagishi, H., Kameshima, S., Schug, J., Koizumi, M., Yamagishi, T., 2010. Theory of mind enhances preferences for fairness 105, 130–137.

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.

Warneken, F., Tomasello, M., 2006. Altruistic helping in human infants and young chimpanzees 311, 1301–1303.

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